UNIVERSITY OF CALIFORNIA SAN DIEGO

Agency and Reflection: Toward an Empirically Adequate Account of Practical Reason

A dissertation submitted in partial satisfaction of the requirements for the degree
Doctor of Philosophy

in

Philosophy

by

Matthew Maurice Braich

Committee in charge:

Professor David O. Brink, Chair
Professor Jonathan Cohen
Professor Matthew Fulkerson
Professor Dana Kay Nelkin
Professor Piotr Winkielman

2020

The Dissertation of Matthew Maurice Braich is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2020

DEDICATION

To my parents, Ted and Lauri Braich, my sister, Kelly Braich, and my wife, Amy Berg.

TABLE OF CONTENTS

LIST OF FIGURES

VITA

2008	Bachelor of Arts, Lewis & Clark College
2011	Master of Arts, University of California, Riverside
2020	Doctor of Philosophy, University of California San Diego

ABSTRACT OF THE DISSERTATION

Agency and Reflection: Toward an Empirically Adequate Account of Practical Reason

by

Matthew Maurice Braich

Doctor of Philosophy in Philosophy

University of California San Diego, 2020

Professor David Brink, Chair

Recent research in social psychology suggests that our attitudes and actions don't typically issue from our reflective capacities. Instead, it appears that they often issue from what psychologists refer to as "automatic" processes, understood as psychological processes that operate quickly, efficiently, and outside of our conscious awareness and control. Much of this same research also suggests that when we do exercises our reflective capacities (rare as it may be), it's usually not to question our

attitudes and actions but rather to rationalize them (that is, to come up with reasons to do, think, and feel things that we want to do, think, and feel anyway). In light of these findings, a number of people have recently argued that we aren't the kind of rational, reflective creatures that philosophers have traditionally thought we are. My aim in this dissertation is to respond to their worries. In the first half, I focus on questions related to our ability to do things for reasons. Common sense suggest that we often do things for reasons, but if our attitudes and actions typically issue from automatic processes, to what extent is this true? I argue that the best way to make good on the idea that we often do things for reasons is to accept both that our ability to do things for reasons doesn't depend on our reflective capacities and that many of the processes that actually guide our attitudes and actions, though automatic, are nevertheless quite responsive to reasons. In the second half of the dissertation, I focus on whether we can rely on our reflective capacities to improve our agency. Although I deny that we need these capacities to do things for reasons, the reasons for which we do things aren't always good reasons, and it's only natural to think that we can rely on our reflective capacities to monitor and control their influences on us. However, this sort of view faces two main empirical challenges. First, it appears that we often don't know what the causes of our attitudes and actions are, so how can we be expected to monitor and control their influences on us? Second, even if we do know what the causes of our attitudes and actions are, why should we think that we're inclined to use reflection to question them? Isn't it much more likely that we'll use it to rationalize them instead? In response to the first challenge, I argue that once we recognize that reflection doesn't

require us to have introspective access to its objects, it becomes more plausible to think that we're often in a good position to reflect on the causes of our attitudes and actions. In response to the second challenge, I argue that as long as we focus on the ways in which other people can improve our reflection, the idea that we'll often be inclined to use reflection to question our attitudes and actions becomes more plausible as well.

Chapter 1

Three Worries about Human Agency

Ordinarily, we think that there are important normative differences between healthy adult humans and other kinds of creatures. It's one thing, for example, if Fido or Baby Charlie were to destroy your couch, but I bet your response would normally be a lot different if it were Uncle Carl instead. But why? What explains these differences? One answer that philosophers have long given is that unlike these other kinds of creatures, healthy adult humans have special kinds of rational capacities: not only can we do things for reasons, but we can also reflect on the reasons for which we do things and try to do things for better reasons if our reasons don't seem very good to us. Indeed, according to this tradition, it's precisely because we have these reflective capacities, whereas young children and non-human animals don't, that we can be held responsible for our behavior.

To see the sort of view that I have in mind, consider a few recent examples.[1] Here's T.M. Scanlon (2000):

> A rational creature is, first of all, a reasoning creature – one that has the capacity to recognize, assess, and be moved by reasons…These reflective capacities set us apart from creatures who, although they can act purposefully, as my cat does when she tries to get into the cabinet where the cat food is kept, cannot raise or answer the question whether a given purpose provides adequate reason for action. We have this capacity, and consequently every action that we take with even a minimum of deliberation about what to do reflects a judgment that a certain reason is worth acting on. (p. 23)

Likewise, Christine Korsgaard (1996) writes that:

> A lower animal's attention is fixed on the world. Its perceptions are its beliefs and its desires are its will. It is engaged in conscious activities, but it is not conscious *of* them. That is, they are not the objects of its attention. But we human animals

[1] For a discussion of the history of this tradition, see Brink (2003; pp. 21-22).

> turn our attention on to our perceptions and desires themselves, on to our own mental activities, and we are conscious *of* them…And this sets us a problem no other animal has. It is the problem of the normative. For our capacity to turn our attention on to our own mental activity is also a capacity to distance ourselves from them, to call them into question…I desire and I find myself with a powerful impulse to act. But I back up and bring that impulse into view and then I have a certain distance. Now the impulse doesn't dominate me and now I have a problem. Shall I act? Is this desire really a *reason* to act? The reflective mind cannot settle for perception and desire, not just as such. It needs a *reason*. Otherwise, at least as long as it reflects, it cannot commit itself or go forward. (pp. 92-93)

And David Velleman (2000) asks us to suppose that:

> [Y]ou were charged with the task of designing an autonomous agent, given the design for a mere subject of motivation…you face a world already populated with lower animals, which are capable of motivated activity, and your task it to introduce an autonomous agent…[Y]ou would add practical reason to the existing design for motivated creatures, and you would add it in the form of a mechanism modifying the motivational forces already at work. You would design practical reason to survey a creature's motives, to block or inhibit some of them, and to reinforce others.
>
> A creature endowed with such a mechanism would reflect on forces within him that were already capable of producing behavior by themselves, as they do in non-autonomous creatures or in his own non-autonomous behavior. His practical reason would be a process of assessing these springs of action and intervening on their operations – which interventions would require an additional, rational spring of action capable of modifying or redirecting the force exerted by the other springs. (pp. 11-12)

I could go on, but I take it that these quotations illustrate the point. For lots of philosophers, human agency depends in some way on our *reflective capacities*. By this, what I have in mind are the kinds of capacities that allow us to do things like think explicitly about our own attitudes and actions, to formulate reasons for those attitudes and actions, and to weigh those reasons in consciousness so that we can arrive at judgments about what our attitudes and actions ought to

be, all things considered.[2] When we exercise these capacities, I'll say that we engage in reflection. So, according to this tradition, exercises of human agency somehow involve reflection. Since it'll be useful to have a name, I'll refer to these "reflectivists views of human agency" (or "reflectivists" views for short).[3]

Reflectivism might be alive today, but is it doing well? Are we really the kind of rational, reflective creatures that reflectivists seem to think we are? Maybe we are, but as a number of people have recently pointed out, there's plenty of research in psychology that seems to suggest otherwise.[4] In particular, if we take this research seriously, then we're likely to think that we don't actually exercise our reflective capacities very often.[5] Instead, it appears that most of our attitudes and actions issue from what psychologists often refer to as "automatic" processes, understood as psychological processes that operate (perhaps among other things) quickly, efficiently, and below the level of our conscious awareness and control. What's more, much of this research also suggests that when we do use our reflective capacities (rare as that may be), it's typically not to *question* our attitudes and actions, but rather to *rationalize* them (that is, to come up with reasons to think, feel, or do things that we're motivated to think, feel, and do anyway). So, in light of these and other related findings, a number of philosophers and psychologists alike have recently argued that we should to reject reflectivists assumptions about human agency. On

[2] The kind of capacities that I have in mind here are similar to what Siegel (2017) calls "upper-echelon rational capacities," which, as she puts it, "include the capacities to deliberate about what to believe, to formulate one's reasons for what one believes or decides, to revise or adjust one's conclusions in light of reflection, including reflection on other things one believes, and on one's reasons for holding one's beliefs" (p. 20).

[3] I borrow this name from Doris (2015).

[4] See, e.g., Haidt (2001), Nahmias (2007), Kornblith (2012), and Doris (2015).

[5] I say "if we should take this research seriously" because research in social psychology has recently come under a lot of criticism. In particular, many classic studies have failed to replicate, and it's been argued when we take this into consideration along with various biases associated with publishing in scientific journals (e.g., the fact that people often don't publish null findings), it starts to look like we have good reason to doubt the reliability of this research. See, e.g., Maxwell, Lau, & Howard (2015). I'll say about more about this "replication crisis" in §3, but I wanted to note it from the outset.

their view, we aren't the kind of rational, reflective creatures that people like Scanlon, Korsgaard, and Velleman seem to think we are.

My aim in this dissertation is fairly straightforward. I want to defend reflectivism against these concerns. Ultimately, what I hope to show is that our reflective capacities can and should play an important role in our practical lives. It's just that research in social psychology shows us that reflectivists need to be more realistic (or at least clearer) about what that role is and how it's performed.

But before I elaborate on this view, I want to take a closer look at some of the relevant empirical research and the worries that it has been thought to raise for reflectivists. In general, I find it helpful to put these worries into three broad groups. The first group concerns the nature of the processes behind most of our attitudes and actions: if most of our attitudes and actions issue from automatic processes, then how often do we actually do things for reasons? I'll call this *the Problem of Automaticity*. The second group of worries concerns our self-knowledge (of lack thereof): given that automatic processes are widely thought to operate below the level of our conscious awareness, to what extent are we aware of the processes behind our attitudes and actions? And if we're not aware of those processes, how can we be expected to reflect on them? I'll call this the *Problem of Self-Knowledge.* The last group of worries has to do with the function or reliability of our reflective capacities: even if we know what the processes behind our attitudes and actions are, why should we think that we'll often use reflection to regulate, as opposed to rationalize, them? I'll call this the *Problem of Unreliable Reflection.*[6] In what follows, I'll try to

[6] Omitted from my discussion are findings in neuroscience that have been taken to show that we lack free will. See, e.g., Wagner (2002). For what I take to be convincing responses to these problems, see Nahmias (2015) and Mele (2014).

motive each of these problems. After that, I'll give a quick preview of how I think reflectivists should respond.

1. Automaticity

Over the last few decades, there's been a massive amount of research in psychology that suggests that automatic processes play a pervasive role in our mental lives. More specifically, what this research seems to show is that our attitudes and actions are typically driven not by the kinds of processes that reflectivists associate with human agency, but instead by processes that are often thought to be (perhaps *inter alia*) fast, efficient, unconscious, and difficult to control.[7] As John Bargh and Tanya Chartrand (1999) put it, "most of a person's everyday life is determined not by their conscious intentions and deliberate choices but by mental processes that are put into motion by features of the environment and that operate outside of our conscious awareness and guidance" (p. 462).

So, what's the evidence for thinking that these processes play such a pervasive role in our mental lives? A lot of it comes from research on priming effects. In general, what this research suggests is that, whether we realize it or not, really subtle and seemingly irrelevant stimuli can have significant effects on our attitudes and actions. Just to name a few examples, studies have found that:

[7] Note that there are a few difficulties concerning how to understand automatic processes. First, the kinds of properties that are associated with these processes can vary from author to author. The characteristics that I've listed above are ones that are perhaps most commonly associated with automatic processes, but they're by no means the only such properties (for a list of many other characteristic that these processes have been said to have, see Frankish and Evans, 2009). Second, even if we focus on the four characteristics mentioned above, it's not clear which (if any) of them are supposed to be necessary. Indeed, as Bargh (1994) points out, many paradigm examples of automatic processes seem to have only some of the relevant properties. To use just one of his examples, driving on "autopilot" doesn't require much attention or effort, and yet the processes involved might be controllable in the sense that driving is typically something that we do intentionally and we can stop these processes at will once they've been set in motion (in this way, they're not like reflexes, which are difficult to stop once they're triggered). So, there are real questions about how we should understand automatic processes. Still, I don't think that these issues affect anything that I say above.

- *Repeatedly exposing people to a stimulus can increase their liking of it.* The classic demonstration of this (also known as the "mere exposure" effect) comes from Robert Zajonc and William Kunst-Wilson (1980). In that study, they subliminally exposed people to randomly generated shapes and then later asked them to evaluate a number of different shapes, only some of which were the shapes to which they were already exposed. What Zajonc and Kunst-Wilson ultimately found was that although many of the participants didn't recognize any of the shapes that they were asked to evaluate, they nevertheless had a strong preference for the old ones, suggesting that the mere fact that they were repeatedly exposed to these shapes had an influence their preferences. Since this study came out, psychologists have conducted hundreds of related studies, using a variety of different stimuli and methods of exposure, and many of them have replicated Zajonc and Kunst-Wilson's basic finding (Bornstein, 1989). What's more, some of these studies have shown that this effect is most likely to obtain when the exposure takes place unconsciously (Bornstein & D'Agostino, 1992).
- *People tend to favor things that they associate with themselves*. For example, Brett Pelham and his colleagues (2002) found that people are more likely to move to states whose name resembles their own (e.g., people named "Virginia" or "Georgia" are 36% more likely to live in Virginia or Georgia). They also found the same effect in the context of job selection (e.g., "Geoffrey"s and "George"s are 42% more likely to be geoscientists). Similarly, other studies have shown that people tend to favor the letters found in their own name (Kitayama & Karasawa, 1997) and that they are more likely to cooperate with others when they have the same birthday (Miller, Downs, & Prentice, 1998). In general, what these and many other studies have been taken to show is that we

tend to be implicit egoists: whether or not we realize it, if we associate something with ourselves, there's a decent chance that we'll like it.

- *The way in which our options are framed can influence our preferences*. For example, Irwin Levin and Gary Gaeth (1988) found that people tend to prefer beef that's described as being 90% lean over beefs that's described as being 10% fat. Likewise, Amos Tversky and Daniel Kahneman (1981) famously found that people are more likely to support a health initiative when it's framed as saving exactly 400 out of 500 lives than when it's framed as resulting in exactly 100 deaths. In either case, because the relevant frames are logically equivalent, it's natural to think that they shouldn't have an influence on people's preferences, and yet these and many other similar studies have repeatedly shown that they do.
- *Seemingly arbitrary stimuli can have significant effects on moral judgments and behavior.* Perhaps the most striking example of this comes from studies on incidental disgust. In general, what these studies suggest is that disgust can amplify our moral judgments even when the source of our disgust has nothing to do with (or is "incidental" to) to target of our judgment. For example, studies have shown that people's moral judgments tend to be harsher when they're sitting at a dirty desk (Schnall *et al.,* 2008), when they smell fart spray (Schnall *et al., 2008*), or when they're under a post-hypnotic suggestion to feel disgust in response to otherwise neutral words like "takes" or "often" (Wheatley & Haidt, 2005). Similarly, studies have shown that self-ascriptions of value can vary depending on really subtle stimuli. For example, Gardner and his colleagues (1999) found that people who were asked to read a story that was told using only first personal *plural* pronouns ("we," "us," "ours") were likely to identify with collectivist

values (e.g. friendship, belongingness, and family), whereas people who were asked to read a story that was told using only first personal *singular* pronouns ("I," "me," "mine") were more likely to identify with individualist values (e.g., autonomy, freedom, and self-sufficiency). Finally, several studies have also shown that exposing people to an image of eyes is likely to elicit pro-social behavior in economic games (e.g., Haley & Fesser, 2005), and Bateson and her colleagues (2006) have more recently found that the same effect obtains in naturalistic settings.

- *Implicit biases can have pernicious effects on our attitudes and actions across a wide range of domains*. For example, Bertrand and Mullainathan (2004) found that the call-back rate for applications with "very African American sounding" names was about 50% lower than it was for applications with "very white sounding" names, even though the applications were the same in all other respects. In another study, Payne (2001) found that priming people with a picture of a Black face made them more likely to mistake an object for a gun. Similarly, studies have shown that in virtual reality settings people (including police officers) are more likely to shoot at an unarmed Black man than an unarmed White man and less likely to shoot at an armed White man than an armed Black man (Correll *et al.,* 2007).

I mention these studies only to illustrate the kinds of effects that are ubiquitous in psychology these days. Indeed, the literature on priming effects is enormous, and there are countless other findings that I could have cited as well. Still, even this small sample is, I think, enough to start to see the problem that this research has been thought to raise for reflectivists conception of human agency. After all, if human agency really does depend on our reflective capacities, as reflectivists

say it does, then shouldn't we expect those capacities to play a much more prominent role in our everyday lives?

To be sure, I don't think that this research raises just one problem for reflectivists. Rather, it seems to raise a variety of problems, depending on what exactly reflectivists are committed to. But in its simplest form, the problem here is often thought to be that the sort of processes that actually issue in our attitudes and actions don't seem to have much in common with the kind of processes that reflectivists associate with human agency.[8] So, If reflectivists are committed to the claim that we often do things based on some kind of reflection, this research would seem to show that they're wrong.[9]

But I'm not convinced that this is the biggest problem that research on automaticity raises for reflectivists. That's because, as I understand it, reflectivism is essentially the claim that human agency depends on our reflective capacities, but of course there are many different ways to understand this dependence relationship. Thus, reflectivists don't have to think that an action is an exercise of human agency just in case it issues directly from the agent's reflective capacities. Instead, they could think that reflection only has to play some kind of indirect role. For example, they could claim that an action is an exercise of human agency as long as it issues

[8] Often this idea is cashed out in terms of a so-called "dual-process" or "dual-system" model of the mind. Very roughly, according to these models, our minds (or only some parts of our minds) involve two different kinds of systems or ways of processing information, one that's usually thought to have many of the features associated with automatic processes and another that's thought to have the opposite set of features – that is, it's slower, less efficient, conscious, and controlled. In recent years, however, these models have become increasingly controversial, and in part that's because many of the operations of our minds resist such a neat categorization. For example, psychologists usually include in the category of "automatic" processes both the intuitive decision-making that guide experts and the really low-level calculations that our visual systems carry out in order to determine the size and locations of objects. However, aside from the fact that these processes are both fast and unconscious, it's hard to see what they have in common. So why stop at just two types of processes or systems? Still, I don't think any of the worries that research in social psychology raise for reflectivists views of agency depend on these models. That's because it's the studies on which these models are based that raise the problems. For this reason, I've decided to eschew talk of these theories altogether.

[9] Haidt (2001) expresses a version of this worry.

from mechanisms that the agent has acquired or developed through reflection or as long as it issues from mechanisms that the agent would endorse on reflection were she aware of them. In either case, since human agency is still being defined at least in part in terms of our reflective capacities, both of these views would seem to count as versions of reflectivism. Yet they don't entail that human agency always requires us to do things directly on the basis of some kind of reflection. So, it's not at all obvious that the main worry that this research raises is that it shows that reflective capacities play only a limited role in our everyday lives. That's a claim that many reflectivists can (and I think *are*) willing to accept.[10]

But with that said, I still think that research on automaticity raises problems for reflectivists views. It's just that the problems that it raises have less to do with the fact that we don't often exercise our reflective capacities and are more directly related to the fact that automatic processes themselves are widely thought to have specific features that seem to be incompatible with certain reflectivists assumptions about human agency. There are three features in particular that I have in mind.

First, automatic processes are widely thought to operate *unconsciously*. Although this can mean different things, what psychologically typically have in mind here is that we often aren't aware of the inputs into these processes, their outputs, or the relationship between their inputs and outputs, where our awareness of something can be understood (or at least measured) in terms of our ability to report on it. For example, in many of the studies above, it's highly likely that, if they were asked, most of the participants wouldn't have been able report on the relevant stimuli or the relationship between the stimuli and their responses. In this sense, the processes behind

[10] See, e.g., Sauer (2012).

their attitudes and actions (and maybe even their attitudes and actions themselves) can be said to be unconscious.

Second, when people are asked about the causes of their attitudes and actions, rather than admitting that they don't know what they are, they will often "confabulate," that is, they will often cite considerations that provide some justification for their attitudes and actions but that played no role in their production. What's more, when people do this, it doesn't appear that they are lying; indeed, studies have shown that people are often highly confident in the accuracy of their confabulations.[11] Nevertheless, under controlled experimental settings, they can easily be shown to be mistaken.

Finally, it also appears that automatic processes can be (and perhaps often are) triggered by *arbitrary* stimuli, meaning stimuli that the participants would not recognize as good reasons for the attitudes or actions that they influence were they aware of them (Doris, 2015). For example, it's hard to imagine that anyone would take the fact that they were exposed to an image of eyes to be a good reason to engage in pro-social behavior or the fact that they are sitting at a dirty desk to be a good reason to make harsher moral judgements. As a result, automatic processes are often thought to be triggered to arbitrary features of our environment and thus "deeply unintelligent" (Stanovich, 2004; p. 34).

If all of this is right, then I think it raises serious questions about whether we really are the kind of rational, reflective creatures that reflectivists seem to think that we are. For one thing, given that automatic processes operate unconsciously, to what extent we are able to engage in the kind of self-reflection that reflectivists associate with human agency? That would seem to

[11] Some of the earliest research on confabulation focused on clinical populations, but it's now pretty clear that health participants are prone to confabulate as well. For a nice overview of this research, see Doris (2015).

require us to have some awareness of the processes behind our attitudes and actions, including some awareness of the inputs into those processes. And yet it's precisely this sort of self-knowledge that we often seem to lack.

For another, if automatic processes are as sensitive to arbitrary stimuli as studies like the ones above suggest, how often do we actually do things for reasons? Granted, we might think that we often do things for reasons, but that could be because we're prone to confabulate. So, despite what reflectivists seems to think, maybe we aren't very rational after all – maybe we don't often do things for reasons.

Just to be clear, reflectivists aren't the only ones who seem to think that we often do things for reasons. In fact, I think it's safe to say that this is a bedrock assumption in most philosophical discussions of issues in ethics and moral psychology, regardless of whether the person making it counts as a reflectivists.[12] So, I take it that concerns about the frequency with which we do things for reasons have a broad scope. They call into question not just reflectivism, but most philosophical views of human agency. Still, since reflectivists tend to locate our ability to do things for reasons in relatively sophisticated kinds of operations, they seem to be particularly vulnerable to these problems.

2. Motivated Reasoning

A natural way for reflectivists to respond to these worries is to argue that it's precisely because automatic processes are so irrational that we need to monitor and control their influence on us.[13] Thus, according to this view, even if our reflective capacities aren't responsible for most of our attitudes and actions, they can still play an important role in our everyday lives, and in

[12] See, e.g., Fischer and Rivizza

[13] Many prominent psychologists hold a view like this as well. See, e.g., Kahneman (2001).

particular we still can use them to make sure that the reasons for which we do things really are good reasons.

However, this view faces its own empirical challenges. One of those challenges should, by now, sound familiar: if automatic processes are unconscious, how can we use reflection to regulate them? But even if we set this worry aside, this sort of view still faces other problems. In particular, there's a considerable amount of evidence that suggests not only that we often don't want to use reflection in this way, but that when we do it's apt to lead us astray. So, to what extent can we count on our reflective capacities to monitor and control the processes behind our attitudes and actions?

As I see it, there are two separate sets of worries here. First, reflectivists often seem to assume people are typically inclined to question their own attitudes actions – that as long as we know what our attitudes or actions are, we'll want to know whether they're reasonable. However, many studies seem to suggest that this isn't the case. On the contrary, it appears that people are much more likely to use reflection to rationalize their attitudes and actions. So, to what extent are we willing to engage in the kind of *critical* reflection that reflectivists associate with human agency?

To see the problem here, consider research on the so-called "my side" bias. In general, what this research shows is that people have a selective tendency to seek out, recall, and accept evidence that supports their views and to ignore, forget, and discount evidence that goes against them. For example, in a classic demonstration of this bias, Lord, Ross, and Lepper (1979) found that people tend to readily accept evidence that's used to support a view with which they already agree. However, when a similar body of evidence is used to support a view with which people

disagree, they'll spend a considerable amount of time and energy trying to come up with reasons to reject it.

Perhaps, then, we shouldn't be too sanguine about the likelihood that people will use reflection to question their own attitudes and actions. Of course, that's not to say that they will never engage in critical reflection, and it's reasonable to think that there will be a decent amount of individual variation here (indeed, philosophers might be an especially reflective bunch). But as a general rule, it seems like people are pretty reluctant to think critically about their attitudes and actions. So, once again, why should we expect reflection to play a regulative role in our everyday lives?

The other set of worries that reflectivists face here has to do with the reliability of our reflective capacities. For the sake of argument, suppose that most people are inclined to question their attitudes and actions. Does that mean that they'll be poised to use reflection to improve their agency? Unfortunately, it doesn't, and that's because human reasoning is subject to a host of other biases as well. In many cases, motivation is again to blame. Indeed, studies have shown that it can bias not only when we reason, but also what we consider while we reason, how much weight we assign to those considerations, and the inferences we draw from them. As Ziva Kunda (1999) puts the point:

> [D]espite our best efforts to be objective and rational, motivation may nevertheless color our judgment because the process of justification construction can itself be biased by our goals…[W]hen we attempt to determine whether a hypothesis is correct, our search for relevant memories and beliefs may be one-sided and may be biased toward finding support for them…This may be all the more true when we are motivated to believe the hypothesis. (p. 224)

For example, studies have shown that it's easier for people to recall experiences in which they acted like an introvert or an extrovert when they're led to believe that being one or the other is desirable (Sanitioso, Kunda, & Fung, 1990). Similarly, research on statistical reasoning has shown that people are more likely to generalize from small samples or to infer causation from mere correlation when doing so supports views that they're motivated to believe (Koosje, Spears, & Koomen, 1995). To make matters even worse, research on "cognitive ease" also suggests that the easier it is for us to recall something, the more likely we are to believe it (Kahneman, 2011). So, it's not just that motivation can bias our reasoning by making it easier for us to recall motivation-congruent information. It's that we're more likely to believe that information because it's easy to recall.

In all of these examples (as well as in many other examples (see Kunda, (1990)), it seems like the participants' motivation to form or maintain certain beliefs is what's driving their biases. But it's important to note that there are many other examples of cognitive biases and errors that don't seem to implicate motivation at all. Confirmation bias is a good example. People don't always selectively seek out evidence that confirms hypotheses because they want to believe those hypotheses. Rather, this seems to be a strategy that we use to test hypotheses in general, whether or not we're particularly motivated to believe them (Nickerson, 1998).[14] Likewise, the heuristics and biases literature contains a litany of "heuristics" or simple rules of thumb that we often rely on when we're making probabilistic judgments, but many of them don't seem to depend on motivation. For example, according to research on the "representativeness" heuristic, people have a tendency to categorize things based on how closely they resemble paradigmatic members of the relevant category (Kahneman & Tversky, 1962). This isn't necessarily

[14] Nickerson, Raymond, 1998, "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology*, Vol. 2, No. 2, pp. 175-220.

something that we do because we're motivated to believe the targets of our judgments belong to the relevant categories. Rather, it seems like it's something that we do to save ourselves time and energy (Kahneman, 2011). Indeed, in many contexts, and especially when the resemblance is diagnostic of category membership, this heuristic can be fairly trustworthy (after all, if something walks and talks like a duck, it's probably a duck). However, the trouble is that we also use this heuristic in situations where the resemblance isn't diagnostic of category membership, and in those cases it's apt to lead us astray.

Moreover, research on conscious versus unconscious decision-making also seems to show that reflecting on our options can lead us to make worse decisions than we otherwise would have made. For example, Wilson and Schooler (1991) asked participants to pick out a poster for their rooms. Half of the participants were told to trust their guts, whereas the other half were given explicit instructions to reflect on their options before making a choice. A few days later, they interviewed the participants to see if they were happy with their choices, and it turns out that ones in the reflection condition were much more likely to be dissatisfied (for similar studies, see also Dijksterhuis, 2004).

If all of this is right, then I think it's reasonable to doubt whether we can really count on reflection to improve our agency. In fact, some psychologists, including Haidt (2001) and Mercier and Sperber (2011), have even taken this research to show that human reasoning didn't evolve to serve an epistemic function (i.e., to help us arrive at true beliefs in some domain or other). Rather, on their views, it serves some other kind of function. For Haidt, reasoning (or specifically moral reasoning) evolved to help us maintain a good reputation. For Sperber and Mercier, it evolved to help us construct and evaluate arguments. Either way, you might think reflectivists are simply wrong about the nature of reflection. After all, they seem to think that we

can count on it to improve our agency, but that might not be what reflection was designed to do, and it might not be very good at it anyway.

3. Shallow Responses

So, is reflectivism doomed? I don't think so, but before I say how I think reflectivists should respond to these worries, I want to consider two other responses. First, reflectivists could simply deny that what they're offering us is a *descriptive* account of human agency. Instead, according to this view, the claim is primarily *normative* – that people should exercise reflective agency, whether or not they do. So, even if we often fall short of reflectivists ideals, it hardly follows that they're wrong.

However, it might be argued this response ignores studies that show that reasoning is apt to lead us astray. For example, if the research on conscious versus unconscious decision-making is right, then it could be that we're often better off not engaging in reflection. So, even when we understand reflectivism in normative terms, as the claim about what we should do, it's still not clear that it's true.

Still, I take it that reflectivists could respond to this worry by arguing that what this research shows isn't that we should abandon reflectivist ideals altogether, but that we need to be more careful with how we go about realizing those ideals (and indeed this would be one area where research on human reasoning could be a good guide for reflectivists, since it could be used to identify the conditions under which reflection is apt to lead us astray). So, at the very least it seems like there's a genuine debate here. If people aren't very good at living up to our ideals, maybe that means we should get new ideals, but it could also mean that we should change how we try to realize them.

Nevertheless, this view strikes me as being too concessive. After all, before we accept that we often don't exercise the sort of agency that reflectivists associate with human agency, we need to answer a host of questions about both what this sort of agency requires and what the science here actually shows. For example, what does it take to do something for a reason, and are automatic processes as irrational as many people seem to think? And do we have to learn about the objects of reflection in any particular way, and to what extent are we truly ignorant of the processes behind our attitudes and actions? In no way are the answers to these questions obvious, and yet they bear directly on issues concerning the frequency with which we exercise human agency. So, I take it that we should try to figure out the answers to these questions first before we concede that we only rarely exercise the sort of agency that reflectivists associate with human agency.

Another way that reflectivists might try to respond to these worries is simply to question the science on which they're based. In recent years, social psychology, and in particular research on priming effects, has come under a lot of criticism, and part of the trouble here is that some classic examples of these effects have failed to replicate. What's more, once we consider the fact that psychologists will often file away studies that don't result in statistically significant effects and that the standards of statistical significance that they use are themselves easy to manipulate, we might start to question the size of these effects, if not their reality as well. None of this, of course, is good news for social psychologists, but it could a blessing for reflectivists. After all, if the science on which the worries above are based isn't very good, then neither are those worries themselves.

However, as John Doris (2015) has recently argued, there are some reasons to think that things aren't as bad in social psychology as they might seem. First, although many studies on

priming effects haven't been directly replicated, they have been *conceptually* replicated, meaning that many other studies have used similar methods to uncover similar effects, even though they aren't designed to recreate the same exact conditions of the studies on which they're based. Second, Doris also notes that some recent meta-analyses of research on priming effects give us reason to be more optimistic about this literature. For example, DeCester and Claypool (2004) looked at over 60 studies that involve at least one of three different kinds of priming effects on impression formation and found that "[t]he mean effect size was significant in each case" (p. 2). Similarly, after reviewing 167 studies on the impact of race-related primes on our attitudes and actions, Cameron, Brown-Iannuzzi, and Payne (2012) conclude that "priming tasks were significantly associated with behavioral measures (r = .28) and with explicit attitude measures (r = .20)" and that they "continued to predict behavior after controlling for the effects of explicit attitudes" (p. 1). So, given these studies and how much conceptual replication there has been in social psychology over the years, Doris argues that we can be fairly confident that priming effects are real, although he also admits that their sizes might not be as large as they are sometimes advertised.

Since I'm not an expert on these issues, I'm happy to take a wait-and-see approach to this debate. Maybe Doris is right, and research in social psychology is more trustworthy than its critics have claimed; or maybe the critics are right, and we shouldn't take this research very seriously – I'll just stay tuned. But in the meantime, reflectivists might want to figure out how to respond directly to the worries that this research raises, or else they're making a pretty big empirical bet on its critics. So, at this point, I think it's safest to operate on the assumption that research in social psychology is generally reliable. In that case, we're right back where we

started: are reflectivists wrong about the nature of human cognition and behavior, or can their view be reconciled with this research?

4. Looking Ahead

For the rest of this dissertation, my task will be to try to answer this question. I'll begin, in chapters 2 and 3, with questions about our ability to do things for reasons. As a background assumption, I take it that most people believe (or at least want to believe) that we often do things for reasons. So, the question that will be driving the discussion here is: how can we vindicate the idea that we often do things for reasons when most of our attitude and actions issue from automatic processes?

In Chapter 2, I'll focus on some recent attempts by reflectivists to answer this question. In particular, according to the views that I'll consider, we can do something for a reason as long as our actions bear the right kind of indirect relationship to our reflective capacities, where what that relationship is will vary from theory to theory. For example, on one version of this view, the idea is that our ability to do things for reasons only requires us to act on the basis of psychological mechanisms that we've deliberately acquired. On another version of the view, it simply requires us to act on the basis of mechanisms that we would endorse on reflection were we aware of them. Since both of these views appeal to our reflective capacities to understand our ability to do things for reasons, they count as versions of reflectivism, and yet they at least promise to make good on the idea that we often do things for reasons even though most of what we do issues from automatic processes.

Although I think these views represent a clear improvement on views that say our ability to do things for reasons always requires us to think before we act, I'm not convinced that they

work. First, they run into clear extensional problems: there will be times when we do things for reasons, and yet we can't appeal to our reflective capacities to explain why. However, as several other people have noted, there's a deeper problem with these views, which is that they're all susceptible to regress worries. After all, reflection itself is often something that we do for reasons, and so we need an account of how this could be. But if we appeal to yet further acts of reflection to do this, we're only moving the question back a step – how could *those* acts of reflection be done for reasons as well? So, unless it's reflection all the way down, it seems like at some point we're going to have to appeal to non-reflective processes to explain how we can do some things for reasons. But in that case, why not appeal to those processes in the first place and avoid the detour through reflection altogether?

In the Chapter 3, I'll try to make good on this suggestion. In particular, I'll argue that the best way to vindicate the idea that we often do things for reasons is to show that automatic processes themselves are capable of giving rise to rational thoughts and actions, independently of their relationship to our reflective capacities. But doesn't research on automaticity show that these processes are often sensitive to arbitrary features of our environment? It might, but as other people have noted, if we take a broader empirical perspective and don't just focus on research in social psychology but consider recent research on reinforcement learning as well, what we find is that many of the processes that psychologists would count as "automatic" appear to be designed to guide us quickly and intelligently in response to meaningful changes in our environment (that is, in response to *reasons*). So, my view will ultimately turn on two main claims: first, that we can do something for a reason if our behavior issues from reasons-responsive mechanisms, and second, that automatic processes often meet this standard, despite what research in social psychology might suggest.

However, even if I'm right, you might wonder what any of this has to do with reflectivism. Indeed, it seems like most reflectivists are committed to saying that our ability to do things for reasons depends, in one way or another, on our reflective capacities. So, insofar as I think that we should reject this claim in favor of a view that locates our ability to do things for reasons in automatic processes, am I really defending reflectivism here or just arguing for an alternative theory?

The short answer is that it depends on whether or not the ability to do things for reasons is supposed to be what makes us special, at least normatively speaking. If it is, then it seems like reflectivists would be committed to the claim that this ability depends on our reflective capacities. But if this ability isn't what makes us special (i.e., if exercises of human agency require us to do more than to simply do things for reasons), reflectivists could still maintain that human agency depends on our reflective capacities without also thinking that the ability to do things for reasons does as well.

My own view is that humans aren't the only creatures who can do things for reasons. In fact, I think lots of creatures can. Still, it seems plausible to me to think that, because of our reflective capacities, we have other kinds of abilities that these creatures lack, and in particular that we have the ability to exercise what I'll call "reflective control" over our attitudes and actions. By this, what I mean is the sort of control that Korsgaard, Velleman, and Scanlon are describing in the quotations above. It's the sort of control that requires us to take a mental step back, as it were, from the reasons for which we do things and to ask ourselves whether we really ought to act on them. Thus, I take it that as long as reflectivists are willing to distinguish between these two abilities (i.e., between the ability to do things for reasons and the ability to exercise

reflective control), then they don't have to say that our ability to do things for reasons depends on our reflective capacities.

However, like any other version of reflectivism, this view is going to face worries about our apparent lack of self-knowledge and the reliability of reflection. In chapter 4, I'll focus on the former set of issues. As I see it, there are two key questions here. First, does reflection require us to have introspective access to its objects, and second, if it doesn't require such access, it is still plausible to think that we often won't be in a position to reflect on the reasons for which we do things. I'll argue that the answer to both of these questions is "No." That is, I think that we can learn about the objects of reflection in many different ways and that once we recognize this it becomes much more plausible to think that we are often in a good position to reflect on the reasons for which we do things.

In chapter 5, I'll turn to worries about the reliability of reflection. Although I don't think that there is an easy way to avoid these worries, I do think that reflectivists would benefit by stressing the ways in which reflection depends on other people. For example, our everyday interpersonal concerns often provide us with the motivation that we need to think critically about our own attitudes and actions, and studies have shown that, in many different context, group reasoning tends to be better than individual reasoning, especially when the members of the group don't already agree on the issue under discussion. So, in the end, I'll argue that if reflectivists stop thinking of reflective control as an individual achievement and instead focus on the ways in which it can be improved by group dynamics, worries about the reliability of reflection start to lose some of their force.

Chapter 2

Reflection and Rational Guidance

What does it take to do something for a reason? And how often do we exercise this ability in our everyday lives? Many reflectivists hold that our ability to do things for reasons requires us to reflect on the processes behind our attitudes and actions. For example, imagine that I want to buy the new iPhone. I might ask myself whether I should act on this desire. Do I have good reasons for wanting the new iPhone, or have I simply been duped by Apple's slick new marketing campaign? According to many reflectivists, it's not until we ask ourselves these questions and decide that, on reflection, the causes of our actions really are reasonable that we can do things for reasons.[1] So, if I decide that I don't have a good reason to buy a new iPhone, but I act on my desire to buy one anyway, then, according these views, I'm not acting for a reason. But, alternatively, if I decide that this desire is reasonable and I act on it, then I would be acting for a reason. We can think of these as "reflection views of rational guidance" (or "reflection views" for short). In general, what these views say is that our ability to do things for reasons depends on our ability to engage in reflection.

But if research on automaticity is right, it's tempting to think that reflection views entail a kind of *skepticism about rational guidance*. After all, it seems like we often don't reflect on the processes behind our attitudes and actions. So, how could it be that, on these views, we often do things for reasons?

A number of people have recently tried to answer this question by claiming that when we do things for reasons, reflection doesn't always have to be the direct cause of our behavior.

[1] I take it that this is the view that Korsgaard is expressing in the quotation that I cited in the last chapter. For other philosophers who seem to who the view that our ability to do things for reasons requires reflection, Elizabeth Anderson (1995) and Tiberius (2002).

Instead, according to these views, it only has to play some kind of indirect role. For example, it could be that our ability to do things for reasons simply requires us to act on the basis of psychological mechanisms that we have acquired or developed in a reflective way or that we would take to be reasonable were we to reflect on them. In either case, since these views explicitly deny that we always have to reflect on the reasons for which we do things right before we do things for reasons, they can at least in principle show that (1) we often do things for reasons, (2) we often do things based on automatic processes, and (3) our ability do to things for reasons depends on our reflective capacities.

Although I think these views represent a clear improvement on versions of the reflection view that requires us to engage in reflection whenever we do something for a reason, I'm still not convinced that they will work. In particular, I think they face two separate worries. First, as other people have argued, any version of the reflection view, whether or not it says that reflection has to play a direct or indirect role in our ability to do things for reasons, is going to face some kind of regress problem. However, even if this problem can be overcome, I'll argue that these views also face extensional worries: there are cases in which we seem to do things for reasons, and yet reflection views don't have the resources to say why. In the end, what I take these problems to show isn't that reflection views obviously are false but that we have good reasons to look for alternatives.

1. Two Theories of Rational Guidance

However, before we take a closer look at reflection views, it might help to say a few words about what exactly theories of our ability to do things for reasons are theories of. To begin, note that philosophers often distinguish between two kinds of reasons. *Normative reasons*, on the one hand, are the reasons that there are for us to do things; they're considerations that

genuinely count in favor of our attitudes and actions. For example, the fact that kicking you in the shin will cause you pain is, at least under normal conditions, a normative reason not to do it; it's a consideration that genuinely counts in favor of not kicking you in the shins. Thus, if I were to kick you in the shins, I would normally be doing something that I shouldn't do (or at least that I have good reason not to do).

Motivating reasons, by contrast, are the reasons for which we do things, and in particular they're considerations that *we take* to be reasons. Obviously, we'll sometimes be right when we take something to be a reason, and in those cases, we can say that our motivating reasons are normative reasons. For example, if I decide not to kick you in the shin because it will hurt you, then the reason for which I decided not to do this (namely, that kicking you in the shin will cause you pain) is a normative reasons. However, in other cases, they can come apart. Imagine, for example, that I have a sadistic normative outlook, and I'm inclined to take the fact that something would cause you pain to be a reason to do it. In that case, if I kick you in the shin because it will hurt you, then my motivating reason wouldn't be a normative reason (in fact, it's a reason not to kick you in the shin). Similarly, imagine that you take a sip of your drink thinking that it's a gin and tonic when in fact it contains petrol. In that case, since you're simply wrong about the contents of your drink, your motivating reasons (name, that your drink is a gin and tonic) isn't a normative reason.[2]

When we're trying to figure out what it is to do something for a reason, there are at least two different questions that we need answer. First, most people assume that motivating reasons are among the causes of our behavior, but how exactly do they enter the causal story? That is, by what mechanism do they get in there? We can think of these as questions about the *informational*

[2] I borrow this case from Bernard Williams (1979)

or *upstream* aspects of rational guidance. Second, when we do something for a reason, it can't just be that our reasons (or our representation of our reasons) causes us to do it. Rather, they have to cause our actions in the right way. But what is the "right" way? What exactly does this relationship involve? We can think of these as questions about the *volitional* or *downstream* aspects of rational guidance.[3]

My focus in this chapter and the next will be squarely on the upstream aspects of rational guidance. This isn't because I think questions about the downstream aspects aren't worth discussing. Rather, it's because research on automaticity raises questions about what it is to take something to be a reason and how often we do things because we take ourselves to have reasons to do them. As I understand them, those are simply question about the upstream aspects of rational guidance.

So, what is it to take something to be a reason? Philosophers tend to have one of two different views about this. According to the first family of views, which I'll call Recognition Views, our ability to do things for reasons requires us to *recognize* things as reasons, where to recognize something as a reason is usually thought to involve some kind of person-level representation of a reason, like a normative belief or judgement. What's more, on these views,

[3] There are at least three different issues that come up in this context. One of them is how to account for cases of "deviant" causal chains. These are cases where an agent is caused to do something by (her representation of) a reason, but the causation happens in such a weird way that it doesn't seem right to say she did it for that reason (Davidson, 1973). So, if reasons are causes, they have to cause us to do things in the right way, but what exactly is the right way? Turns out, it's really hard to answer this question (for one attempt to do this, see Mele (1987); for a recent criticism of Mele, see Sehon (2017)). Another issue that comes up in this context concerns the nature of normative motivation: what kind of mental states does this species of motivation involve? For example, does it always involve beliefs and desires (as, e.g., Smith (1994) argues), or in some cases are normative beliefs capable of motivating us on their own (as, e.g., Nagel (1979) claims)? Finally, the last issue that comes up in this context is whether or not reasons-explanations are causal explanations. Davidson (1963) argued that they are, and most philosophers agree, but there are a few exceptions (see again Sehon (2017)). Nevertheless, I'll assume that reasons-explanations are causal explanations.

when this state of recognizing something as a reason guides our behavior in the right way, we've done something for a reason.

But some philosophers think that the upstream aspects of rational guidance require more than this. On their view, when we take something to be a reason, it's not enough to simply recognize it as such. Rather, we also have to engage in reflection; that is, we have to think explicitly about whether the considerations that we recognize as reasons to do things really are reasons to do them. So, unlike recognition views, which say that reasons enter into explanations of our behavior *via* the personal-level states that represent them, these views say that they first have to go through reflection.

In the next chapter, I'll have more to say about recognition views. But for now, let's set them aside. What I want to focus on in this chapter is the second kind of view, what I earlier called "reflection views."

2. Three Kinds of Reflection Views

The version of the reflection view that's most vulnerable to worries about automaticity are what I'll call *proximate* reflection views. According to these views, whenever we do something for a reason, our behavior has to be directly based on some kind of reflection. Thus, on these views, if our attitudes and actions instead often issue from automatic processes, they can't be had or done for reasons.[4]

However, as I mentioned earlier, there are two other version of the reflection view. One of them is what I'll call *distal* reflection views. In general, these are views that say that we can

[4] Note that this requires the further assumption that an attitude or action can't issue both from an automatic process and a reflective one. This is a simplifying assumption, of course, and most likely false in very many cases, but for the purposes of illustrating the problem that research on automaticity poses for these views, I take it that it's relatively harmless.

do things for reasons as long as reflection plays the right kind of role in the etiology of our behavior. So, according to these views, while our ability to do things for reasons does require us to engage in reflection before we act, it doesn't have to be right before we act. Instead, it can be at an earlier point in our lives.

The third kind of view is what I'll call *possible* reflection views. According to these views, we can do something for a reason as long as our behavior bears the right relationship to some possible act of reflection. Thus, unlike the other two versions of the reflection view, this version doesn't say that our ability to do things for reasons requires us to engage in reflection at all (whether it's right before we act or not). Instead, what matters is what you would have thought had you engaged in reflection.[5]

If we accept either of these views, then the fact that our attitudes and actions often issue from automatic process doesn't by itself show that we don't often do things for reasons. Rather, in order to draw that conclusion, we need to answer two further questions. First, what is the relationship that, according to these theories, our actions must bear to possible or distal acts of reflection? And second, how often do actions that issue from automatic processes stand in that relationship, whatever it is?

To see how these questions could be answered, consider what I take to be the most prominent version of each view. The first is a version of the distal view that Julia Annas (2011) and Hanno Sauer (2012) have recently defended. According to this view, we can do something for a reason as long as our behavior is guided by psychological mechanisms that we've acquired or developed reflectively. To illustrate this view, Annas appeals to various kinds of skills. Early

[5] Arpaly and Schroeder (2009; 2015) distinguish between these three versions of the reflection view as well.

on in the process of acquiring a skill, she tells us, we might often imitate an expert's behavior. But at some point, she writes,

> To acquire the skill, you have to be able to do it yourself, rather than stopping at a plateau of routine, where you can turn off thinking about further improvement. This is a point familiar in a whole range of skills: the moment comes when you have to stop just following the teacher and play, skate, dance, speak in Italian, for yourself. It is also clear that this is the point of the instruction: if you can't do it yourself in a way that is not merely parroting the teacher then you have not yet learned the skill. This point about self-direction links naturally to the first point: you have to make the effort to understand what you have been taught, and to grasp it for yourself, because this is the point at which you can exercise the skill in a self-directed way. (pp. 17-18).

In addition, Annas claims that skill acquisition also involves an "aspiration to improve." "This," she writes, "is what a lot of practice is about – not perfecting a routinized movement but learning to do what is being done, but better: how to steer a car, skate a double axel, translate Homer, clear a hurdle" (p. 18). For example, becoming a better golfer requires you to think about your swing and to make any changes that you feel are necessary by going to the range and practicing. Thus, according to Annas, acquiring a skill – or at least acquiring what she calls a "self-directed" skill – requires us to reflect on the processes behind our behavior and to try to correct them if we see room for improvement. Of course, we make these changes hoping that they'll eventually become "second nature" or automatic (you don't want to think too much about your swing when you're really playing), but that doesn't mean that the behavior in which they issue isn't done for a reason. On the contrary, according to this view, even though these processes are automatic, because we have explicitly trained them in a way that was responsive to reasons, any behavior in which they issue is done for a reason as well. To give it a name, we can think of this as *skill model of rational guidance.*

The other view that I have in mind is a version of the counterfactual view. In general, the thought here is that we can't do things for reasons unless we act on the basis of considerations that, were we to reflect on them, we would recognize them as reasons for the actions that they influence. Note, however, that unlike the skill model, this view doesn't state a sufficient condition of rational guidance. Rather, it only states a necessary condition. That's because there could very well be cases in which our behavior satisfies the counterfactual condition, and yet it isn't done for reasons. For example, I take it that most people would want to say that we don't do things for reasons when we're sleepwalking, and yet it could be that somnambulant behavior is often done in response to considerations that we would recognize as reasons were we to reflect on them.[6] So, anyone who accepts this sort of view would presumably want to say that it serves simply as a constraint on rational guidance.[7] That is, on this view, we can't say for sure whether or not someone did something for a reason just because their behavior satisfies this counterfactual condition, but we can say for sure that they didn't do something for a reason if it doesn't. I'll call this the "possible reflection" view.

To see the difference between these views, just imagine a case in which you do something automatically, but the processes behind your behavior weren't acquired reflectively. Still, it just so happens that, on this occasion, those processes are triggered by a stimulus that you would recognize as a reason were you aware of it. For example, imagine that you're on a hike, and right as you're about to take another step, you look down, see a snake, and automatically jump away. It's probably safe to assume that, on reflection, you would take the fact that you're

[6] See, e.g., Levy and Bayne (2004). I would like to thank Dick Arneson and Saba Bazargan-Forward for pressing me on this point.

[7] For example, Tiberius (2012) says that when an agent does something for a reason "she must take herself to have good reasons for the option she chooses…[and her taking herself to have good reasons] must not change under appropriate reflection" (p. 343). Likewise, Scheffler (1994; p. 30) defends a similar view, and Doris (2015) understand reflectivism along these terms.

about to step on a snake to be a good reason to jump away, and yet it's highly unlikely that you trained yourself to do this – instead this tendency is probably hard-wired or the product of implicit learning processes. As a result, on the skill model, we couldn't say that your behavior isn't done for a reason; to put it plainly, you simply weren't acting on the basis of processes that you acquired reflectively. However, because you were acting in response to considerations that, on reflection, you would recognize as reasons for your behavior, the possible recognition view could, depending on what else it's thought to involve, entail that you did something for a reason.

Since both of these views allow us to say that we can do things automatically and yet still for reasons, I take them to have a clear advantage over views that say that we have to think about what to do right before we do things for reasons. What's more, because they both still explain our ability to do things for reasons at least partly in terms of our ability to engage in reflection, they should also count as versions of the reflection view. So, are we done here? Does either view give reflectivists a satisfying way to avoid the problem of automaticity? Unfortunately, I don't think they do.

3. Reflection All the Way Down

The first problem with all of these views is that they're susceptible to regress worries. To see these worries, first note that reflection is not only something that we do, but also often something that we do *for a reason*. However, if that's true, then we need some way to explain how it is that acts of reflection could be done for reasons when they are. One way to explain this would be to appeal to some kind non-reflective processes that ground our ability to do things for reasons. But because reflection views are committed to the claim that our ability to do things for reasons depends on our reflective capacities, they'll have to explain this by appealing to further acts of reflection. But, of course, those acts of reflection are also often going to be done for

reasons, and so to explain how that could be they'll have to appeal to yet further acts of reflection, and so on, *ad infinitum.*

Peter Railton (2009; m.s.), Hilary Kornblith (2012), and Nomy Arpaly and Timothy Schroeder (2012; 2015) have all recently developed versions of this worry. However, because Arpaly and Schroeder explicitly develop a version of this argument for each of the reflection views mentioned above (that is, for proximate, distal, and counterfactual versions of the view), I'll focus on their formulations of it. As I understand it, their argument turns on two key claims. First, they claim that every action, whether it's overt or mental, is done for a reason. Second, they claim that reflection is always a mental action. Thus, if we accept these two claims, and if we think that our ability to do things for reasons depends in some way on our reflective capacities, then it seems like we're going to encounter an infinite regress. To put it schematically, here's is how I understand their argument:

P1. Every action is done for a reason. (First Assumption)

P2. Reflection is an action. (Second Assumption)

P3. If we perform an action for a reason, then it must bear the right relation to some proximate, distal, or possible act of reflection R. (Reflection Views)

P4. But since R is an action, it too must be done for reasons, which means that it must bear the right relationship to some further proximate, distal, or possible act of reflection R*. (From P1-P3)

P5. But since R* is an action, it too must be done for a reason, which means that it must be the right relationship to some further proximate, distal, or possible act of reflection R**. (From P1-P4)

P6. But since R** is an action, it must be done for a reason, which means that…

Now, the precise details of this argument will vary depending on what the target version of the reflection view is. For example, if we're targeting the proximate view, then argument would be as follow:

P1. Every action is done for a reason. (First Assumption)

P2. Reflection is an action. (Second Assumption)

P3. If we perform an action for a reason, then it must issue directly from an act of reflection R. (Proximate View)

P4. But since R is an action, it too must be done for a reason, which means that it must issue directly from some other act of reflection R*. (From P1-P3).

P5 But since R* is an action, it too much be done for a reason, which means that…

Alternatively, if we focus on the counterfactual view that I discussed above, then the argument would be:

P1. Every action is done for a reason. (First Assumption)

P2. Reflection is an action. (Second Assumption)

P3. If we perform an action for a reason, then it must be done in response to considerations that we would recognize as reasons were we to engage in some act of reflection R. (Possible Reflection)

P4. But since R is an action, it too must be done for a reason, which means that it must done in response to consideration that we would recognize as reasons were we to engage in some further act of reflection R*. (From P1-P3).

P5. But since R* is an action, it too much be done for a reason, which means that…

Finally, since the skill model only states a sufficient condition, it might be thought that it's not susceptible to this argument. After all, in order for the regress argument to work, (P3), the premise that states the relevant version of the reflection view, has to be formulated in terms of a necessary condition. Otherwise, advocates of the relevant view wouldn't be committed to saying that acts of reflection *have to be* related to further acts of reflection, which is what sets the regress in motion. However, I decided to formulate the skill model as a sufficient condition because I assume that anyone who holds this view would also want to say that we can do things for reasons if our behavior issues directly from our reflective capacities as well. So, I take it that this view is best thought of as a disjunctive claim: that if we do something for a reason, our behavior must issue either directly from reflection or from mechanisms that we've acquired reflectively. In that case, since the left-hand side of the disjunct will run into the same sort of regress problem that the proximate view encounters, I'll focus only on right-hand side. Thus, the regress argument against this view would be:

P1. Every action is done for a reason. (First Assumption)

P2. Reflection is an action. (Second Assumption)

P3. If we perform an action for a reason, then it must issue from mechanisms that we acquired through an act of reflection R. (Skill Model)

P4. But since R is an action, it too must be done for a reason, which means that it issue from mechanisms that we acquired through an act of reflection R. act of reflection R*. (From P1-P3).

P5. But since R* is an action, it too much be done for a reason, which means that…

In short, then, since there are many different kinds of reflection views, there will be many different kinds of regress arguments against them. Arpaly and Schroeder provide us with a general scheme for thinking about the form that these arguments take, but it's not too difficult to fill in the details once we have a target view. Thus, this discussion hopefully shows that as long as we hold that rational guidance depends on reflection, that reflection is an action, and that all actions are done for reasons, we'll have to think that our ability to do things for reasons requires infinitely many acts of reflection.

So why is this a problem? Arpaly and Schroeder offer us different answers depending on when or where the relevant acts of reflection have to take place. First, they argue that, because of our limited time and cognitive resources, we couldn't possibly engage in infinitely many acts of actual reflection. So, any view that says that our ability to do things for reasons requires actual reflection, which includes the proximate and the distal forms of the reflection view, is committed to saying that we can't do something for a reason unless we do something that's psychologically impossible.

Second, Arpaly and Schroeder argue that any view that requires us to engage in infinitely many acts of possible reflection is going to face an epistemic problem. Here is how they put the worry:

> Possible Deliberation, like its counterparts, faces a serious regress problem. Suppose it is true that, had Harold deliberated about the reasonableness of calling to mind promises that might conflict with meeting his son in Calgary, he would have reached the (theoretical) conclusion that it would be reasonable of him to do so. But had Harold deliberated he would have done so for a reason; and for this counterfactual act of deliberation to be one that would have been done for a reason, it would have to be true that had Harold deliberated about the counterfactual act of deliberation, he would have come to a conclusion as to how to deliberate about how to deliberate about the reasonableness of calling to mind promises that might conflict with meeting his son in Calgary. That in turn would have had to have been done for a reason, and so on. So Harold's mind, as it actually is, must support infinitely many counterfactuals about the ever increasingly complex processes of reasoning he would have gone through. And Harold, being an ordinary human being, is not capable of guaranteeing that if he had deliberated about how to deliberate about how to deliberate about...about how to deliberate about meeting his son in Calgary, he would have reached a policy about deliberating about deliberating about...If Harold had attempted to engage in this sort of massively complex deliberation, he would have become confused, and given up, and reached no conclusion at all. This is the viciousness of this particular form of regress. (p. 31)

As I understand it, the problem here is that, according to Arpaly and Schroeder, possible reflection views require us to know what the outcome of infinitely many acts of possible reflection would be before we can know whether any of our actions are done for a reason. But of course, given our limited psychological capacities, we couldn't possibly know what the outcome of all of these acts of possible reflection would be. Thus, on these views, we can't know whether our actions are done for reasons.

So, how might people who accept some version of the reflection view respond to these arguments? One possibility would be to accept the regress and to argue that it isn't as vicious as

Arpaly and Schroeder think. Now, I take it that this strategy probably won't work in defense of views that require infinitely many acts of actual reflection (indeed, this version of the regress does strike me as vicious), but it might work in defense of possible reflection views. For example, I don't think it's obvious that these views commit us to thinking that our ability to know whether we do things for reasons requires us to know what the outcome of infinitely many acts of possible reflection would be. Perhaps instead they could say that it simply requires us to know what the outcome of some of these acts of possible reflection would be and to generalize from there. Alternatively, it might also be thought that we should keep metaphysical and epistemic claims separate. So, perhaps it's not incumbent on people who accept the claim that our ability to do things for reasons requires us to engage in infinitely many acts of possible reflection to say how we know whether any of our actions are done for reasons. That would simply require another theory.

Nevertheless, it seems to me that, all else being equal, a theory that doesn't entail an infinite regress is better than one that does. So, I take it that, before they bite the bullet and accept the regress, advocates of the reflection view should try to come up with some way to avoid having to say that our ability to do things for reasons requires us to engage in infinitely many acts of reflection, whether they're actual or merely possible. What's more, I take it that the best way to do this is to reject either (P1) or (P2). So, what I want to consider now is whether we should accept either of these claims and, if not, whether rejecting them would help advocates of the reflection view.

4. Blocking the Regress

Let's start with (P1), the claim that all actions are done for a reason. Arpaly and Schroeder certainly aren't the only philosophers who accept this claim. In fact, there's a prominent tradition

in action theory, going back at least to Anscombe (1957) and Davidson (1963), that says that what separates actions from other things that we do is that only actions are done for reasons.[8] Thus, according this tradition, if you want to know how to distinguish, say, a wink from a blink, you have to look inside the agent's head: winks are things that we do for reasons, whereas blinks are mere reflexes.

But not everyone agrees with this claim. For example, Rosalind Hursthouse (1991) has argued that sometimes when we act on our emotions, we do things intentionally but not for reasons. She calls these "arational actions," and one of her examples involves someone who, while in the grip of anger, suddenly kicks a table. According to Hursthouse, when this person kicks the table, he's acting intentionally, and yet we needn't suppose that there was anything that he took to count in favor of his behavior. Instead, it could be that he kicked the table simply because he was angry. Likewise, Warren Quinn (1994) asks us to imagine someone who turns on radios whenever he sees one not because he thinks there is anything worthwhile about doing this, but out of a compulsion. Again, Quinn's thought seems to be that while this man is acting intentionally, he isn't acting for a reason.

If Hursthouse and Quinn are right, then (P1) is false – not every action is done for a reason.[9] As a result, advocates of the reflection view might try to block the regress by modeling the acts of reflection on which our ability to do things for reasons are based on these arational or compulsive actions. If so, then they could still maintain that our ability to do things for reasons requires us to engage in reflection without having to explain how those acts of reflection are themselves done for reasons.

[8] See Wilson and Shpall (2002) for a nice overview of this tradition.

[9] Of course, this isn't to say that Hursthouse and Quinn are right. See, e.g., Smith (1998) for a response to both criticisms. However, if neither is right, then that would seem to be a problem for people who accept the reflection view. So, if only for the sake of argument, I'm happy to grant that Hursthouse and Quinn are right.

But I am not sure that this strategy will work. For one thing, as Railton (m.s.) points out, it's difficult to imagine that our ability to do things for reasons could depend on such pathological or defective exercises of agency. For another, I doubt that most of us have engaged in arational or compulsive acts of reflection (usually, when I'm upset, for example, I don't find myself suddenly reflecting on the processes behind my attitudes and actions). So, it seems like any version of the reflection view that requires actual reflection is going to encounter familiar worries about the frequency with which we do things for reasons. Thus, I take it that even if (P1) is false, a modified version of the claim still poses a problem for the reflection view. On this modified version, the thought is that either the acts of reflection on which our ability to do things for reasons is based are done for reasons or they're compulsive or arational. If it's the former, then advocates of the reflection view will face a regress. But if it's the latter, then the view start to lose much of its appeal.[10]

However, there could be a third option available to advocates of the reflection view who want to block the regress by arguing that the acts of reflection on which our ability to do things for reasons is based aren't themselves done for a reason. Rather than appealing to compulsive or arational actions, they could instead appeal to cases in which agents have to make what we might think of as "hard choices." More specifically, what I have in mind are cases in which the agent has reason to do A, has reason to do B, but she doesn't more reason to do A instead of B (and *vice versa*). Still, she has to pick one of these options, and so her choice, whatever it ends up

[10] As I read him, Railton (m.s.) formulates his version of the regress argument in these terms. That is, he states it as a dilemma: either the acts of reflection that ground our ability to do things for reasons are themselves done for reasons, in which case the reflection view face a regress, or they're not done for reasons, in which case the reflection view seems unmotivated.

being, won't be based on the reasons that are available to her. Instead, it'll have to be based on something else.[11]

In debates about free will, some philosophers have taken actions that are the result of hard choices to ground our ability to act freely.[12] Could the same thing be said about our ability to act for reasons? Just to be clear, since what we want to know is whether or not the acts of reflection on which our ability to do things for reasons is based have to be done for reasons, those acts of reflection themselves would have to be the objects of our hard choices. Thus, the question here isn't whether our ability to do things for reasons requires us to make some kind of hard choice or other; rather, it's whether it requires us to make hard choices that specifically involve an option to engage in reflection. So, if advocates of the reflection view want to block the regress by appealing to hard choices, what they would really need defend is the claim that our ability to do things for reasons requires us to engage in either actual or merely possible reflection as a result of one of these choices.

But even if this claim allows advocates of the reflection view to block the regress, it comes with some serious costs. Again, hard choices seem to be pretty rare, and this is especially true when we focus on cases in which one of our options is to engage in reflection. So, it's difficult to see how any version of the reflection view that requires us to engage in reflection as the result of hard choices that we've actually made (whether they're right before we do things for reasons or at some much more remote point in the past) could make good on the idea that we often do things for reasons.

[11] For some putative examples of such choices, See Chang (2002)
[12] See Kane (1998).

If that's right, then perhaps the only viable alternative is a version of the possible reflection view that says that our ability to do things for reasons requires us to engage in reflection as the result of a hard choice that we've made in some possible world. Admittedly, I don't have a knock-down argument against such a view, but it does seem quite *ad hoc* to me. After all, why would anyone think that our ability to do things for reason require such a bizarre counterfactual to be true if not only because they want to avoid the regress problem? So, I take it that before advocates of the reflection view go this route, they might first want to consider some other ways to block the regress.

So let's now turn to (P2), the claim that acts of reflection are mental actions. To be more precise, Arpaly and Schroeder's claim here isn't simply that reflection is a mental action, but that it's constituted by a bunch of discrete mental acts that are themselves actions. For example, when we engage in reflection, we have to *bring to mind* reasons for and against our options, *weigh* those reasons against each other, and *arrive at* a conclusion about what the most reasonable option is, among other things. On Arpaly and Schroeder's view, each of these discrete mental acts (i.e., the *bringing to mind*, *weighing*, and *arriving at a conclusion*) are actions and thus, given (P1), done for reasons.

But the claim that all of the mental acts that make up reflection are genuine actions is again controversial. Al Mele (2009), for example, argues that at least some of the mental acts in which reflection consists aren't always actions, and in particular he thinks this is true of *remembering*. As I understand it, his argument is that actions essentially require effort on the agent's part (that is, she must *try* to do something), but remembering is often an effortless activity. Similarly, Galen Strawson (2003) goes one step further and argues that hardly any of the mental acts in which reflection consists are mental actions. This is because, according to

Strawson, they typically occur automatically. Hence, he writes that "very often [when we engage in reflection] there is no action at all: none of the activation of relevant considerations is something one does intentionally. It simply happens, driven by the practical need to make a decision. The play of pros and cons is automatic – and sometimes unstoppable" (p. 243). He then goes on to add that, "the movement of the natural causality of reason to its conclusion in choice or decision is lived (by some) as action when it is really just reflex; distinctively rational reflex, to be sure, but not in any case a matter of action" (p. 244).

Notice, though, that if advocates of the reflection view argue against (P2) on these grounds, they'll find themselves in a precarious position. After all, what we're ultimately trying to figure out is how automatic processes can give rise to rational thoughts and actions – how it is, in other words, that we can still think, feel, and do things even when our attitudes and actions issue from automatic processes. And advocates of the reflection view tell us that this is possible just in case automatic processes bear the right relationship to acts of reflection. But if, in order to avoid the regress problem, they go on to say that acts of reflection are not themselves actions because they too consist in automatic mental activities, then they seem to face a dilemma: either those mental activities are done for reasons, in which case they will only push the problem back a step, or they're not, in which case it seems like we still need an explanation for how it could be that automatic processes could give rise to rational thoughts and actions, only now we're focusing specifically on the automatic processes that give rise to the mental activities in which reflection consists.[13] Of course, advocates of the reflection view could say that when automatic processes are involved in reflection, they acquire the ability to confer rational properties onto other automatic processes, or they could say that the sum of reflection is more rational than its

[13] For a similar argument, see Railton (2009; m.s.).

automatic parts. But without further explanation of how this could be, both of these responses seem mysterious. That is, why would automatic processes be able to give rise to rational thoughts and actions only when they're involved in reflection? What is it about their contributions to reflection specifically that makes them special?

I'm willing to leave open the possibility that advocates of the reflection view can answer these questions. However, I do think they force us to confront what I take to be a central problem for their view: namely, that when psychologists tell us that automatic processes play a pervasive role in our minds, they mean all parts of our minds, including reflection. Indeed, as Mele and Strawson remind us, reflection always seems to involve automatic processes. So, we have a choice: we can either continue to insist that the automatic processes in which reflection consists are somehow special or try to give an account of how these processes can guide rational thoughts and actions independently of their relationship to our reflective capacities. My own view, which I develop in the next section, is that the latter will ultimately give us a better response to the problem of automaticity.

5. Extensional Problems

Nevertheless, let's suppose that advocates of the reflection view can avoid the regress problem – maybe that's because they're willing to tolerate the regress, or maybe it's because they have a way to block it. Does that mean they're in the clear? I don't think so, and that's because I don't think they can account for the full scope of behavior that seems to be automatic and yet still done for reasons.

To see why, let's start with the skill model of rational guidance. Recall that, according to this view, we can do things automatically and yet still for reasons as long as our behavior issues

from mechanisms that we've acquired reflectively. However, I take it that this view really involves two separate claims: that an action can issue from automatic processes and yet still be done for reasons as long as (1) those processes are learned responses to the stimuli in question (call such processes "skills") and (2) those skills were learned as a result of a processes that involved reflection. Now, (2) is what makes this a version of the reflection view, but is it necessary? Take, for example, language acquisition. Many of the processes that guide our linguistic behavior are acquired not through explicit instructions and reflection, but rather though *implicit* learning processes, and yet they're fully capable of guiding our linguistic behavior in intelligent and flexible ways to an open-ended array of stimuli (Reber, 1987; Dienes & Altmann, 1997).[14] So, although (1) is true of these processes, (2) isn't, and as a result we might wonder whether or not it's really needed to explain how automatic processes could give rise to rational thoughts and actions. Might not a claim like (1) be enough?

Of course, I mention our linguistic tendencies just to illustrate the point that some of our behavior is, on the one hand, quite responsive to reasons but, on the other hand, guided by psychological processes that are acquired through implicit learning. But I also think that it's easy to find many other examples of this. For instance, most of us are pretty good at detecting subtle indicators of intent (Frank, Gilovich, & Regan, 1993; Reed, Zeglen, & Schmidt, 2012), and yet it's highly unlikely that this is an ability that we learned explicitly. Or consider our tendency to mimic to tone, mood, posture, and gestures of people with whom we're interacting. Studies have shown that such "mimicry" behavior is quite selective (Chartrand & Lakin, 2012) and that it can have a wide range of benefits (Chartrand & Bargh, 1999; Balinson & Yee, 2005; Stel, van de Bos, & Bal, 2010), but again this isn't something that we learn to do explicitly. Finally, just to

[14] Charlie Kurth (2018) raises a similar problem for Annas's view.

mention one more example, studies have shown that when we're making decisions in situations where we're not sure about what the expected value of all of our options is, we'll often learn implicitly which of them is best well before we can articulate why (Bechera, Damasio, & Trenel, 1997). As with the other examples above, I take it that what we have here is a good candidate for reasons-responsive behavior, but the processes that guide it aren't acquired in the way that the skill model would predict.

If all of this is right, then it seems safe to say that the skill model overintellectualizes rational guidance. That is, although I'm inclined to agree with advocates of the skill model that the key reconciling research on automaticity with the idea that we often do things for reasons is to look at the role that skills play in our everyday behavior, we shouldn't assume that the skills that are implicated in rational guidance have to be acquired explicitly. Thus, I think we should accept (1) but reject (2).

It also seems to me that the version of the possible reflection view that I considered above – namely, the idea that we can't do something for a reason unless we act in response to considerations that we would recognize as reasons were we to reflection of them – faces extensional worries as well. In particular, I take it that while this claim gives us a good test for determining whether or not the considerations on which we act are congruent with our explicit attitudes, it's not a good test for determining whether or not those considerations count as the reasons for which we do things. This is because there are times when we do things for reasons, but we wouldn't recognize them as such on reflection. Thus, it's these kinds of cases (i.e. cases in which our motivating reasons conflict with our explicit attitudes) that pose a problem for this version of the possible recognition.

To see what I have in mind, consider two different scenarios. In the first scenario, you act in response to considerations that genuinely count in favor of your behavior, but because your explicit attitudes are misguided, you wouldn't recognize those considerations as reasons were you to reflect on them. For example, consider the famous Huck Finn case.[15] As the story goes, Huck has a choice to make between turning in his friend Jim, who's a run-away slave, or helping him escape the local authorities. Explicitly, Huck thinks that he ought to turn Jim in, but because they're such good friends, he can't bring himself to do this. So, he ultimately decides to help Jim escape. Now, it seems to me that when Huck decided to help Jim out, he was acting in response to reasons (indeed, he was acting in response to very good reasons), but because he has grown up in a racist society and has internalized many of its norms, we can imagine that he wouldn't recognize them as such on reflection.

In the second kind of scenario that I have in mind, your explicit attitudes aren't misguided, but you act in response to bad reasons anyway. For example, imagine that someone who is committed to egalitarian values but, despite himself, acts on the basis of an implicit bias. In that case, although this person might not recognize the considerations on which he acted as reasons for his behavior on reflection, we might still think that they were motivating reasons (albeit really bad ones). For example, imagine that Andy, a white man with explicit egalitarian values, is at the mall and sees a black man holding a purse. Let's also suppose that, because of his bias, Andy implicitly takes the fact that he sees a black man holding a purse to count in favor

[15] For an extended discussion of this case, see Arpaly (2003). One response that advocates of the reflection view might make is that Huck's emotional response is the product of *implicit reflection*. However, unless the emotions associated with friendship are always the products of such processes, I think we can easily imagine a version of the case in which this isn't true. What's more, I think that there's a genuine question about whether we can even make sense of the notion of implicit reflection, since it's almost always defined as an essentially conscious activity. See, e.g., Haidt (2001). Still, in the next chapter, I'll claim that our intuitive responses are often the products of sophisticated learning mechanisms that select actions in ways that resemble what we would ordinarily call reasoning, but I wouldn't call this reflection, implicit or otherwise.

of thinking that he stole it, and so he says to himself, "I bet he stole that purse." Now, given Andy's explicit attitudes, we can assume that he would say, on reflection, that the fact that a black man is holding a purse is not a reason to believe that he stole it, but it's still plausible to think that this consideration was the reason for his racist thought.

Notice that both of these cases have a common structure. In each, the agent is acting on the basis of considerations that are arguably (i) motivating reasons but that (ii) conflict with their explicit attitudes. But, of course, the cases differ with respect to the quality of the relevant reason and the agent's explicit attitudes. In the first case, the agent's explicit attitudes are bad, but his reasons are good, whereas in the second case, the agent's explicit attitudes are good, but his reasons are bad. Nevertheless, it's this basic structure that generates problems for the possible recognition view. After all, if our ability to do things for reasons requires us to act on the basis of considerations that we would recognize as reasons on reflection, then (ii) would seem to preclude the possibility of (i). So, in other words, according to this view, it shouldn't be possible to do things for reasons that are bad by the agent's own lights, and yet that's precisely what seems to be happening in these cases.

In short, then, neither the skill model nor the version of the possible reflection view discussed above can account for the full range of behavior that issues from automatic processes but is still done for reasons. For the skill model, the difficult cases to explain are cases where we act on the basis of psychological mechanisms that seem be reasons-responsive but that we're acquired explicitly. For the possible reflection view, the difficult cases are cases where we act in response to considerations that seem to be motivating reasons but that we wouldn't recognize as reasons on reflection.

Of course, these are just two versions of the reflection view, and so it's possible that there could be a version of the distal view or the possible reflection view that would allow us to avoid these worries. However, I'm tempted to think that both of these problems generalize. First, it's plausible to think that there will be times when we act on the basis of reasons-responsive psychological mechanisms that we acquired through implicit learning processes and reflection played no role whatsoever in the production of our behavior. And if that's right, then no version of the distal view will be able to explain these cases. Second, it also seems plausible to think that there will be times when we act on the basis of psychological processes that we would completely disavow on reflection, but our behavior is nevertheless done for a reason. And if that's right, then it's hard to see how any version of the possible reflection view will be able to account for these cases as well.

6. Conclusion

In this chapter, I've argued that we have good reasons to doubt whether our ability to do things for reasons depends on our reflective capacities. In the first place, all versions of this view seem to be vulnerable to regress problems. However, even advocates of this view can avoid this problem, they will still encounter other worries. In particular, at least two of the most prominent versions of this view can't account for certain cases in which our behavior issues from automatic processes but is still done for a reason, and I suspect that the worry here might generalize to other versions of the view as well.

Still, I'm willing to admit that nothing that I have said here constitutes a knock-down argument against the reflection view. But I do think the arguments above give us at least some reason to start looking for an alternative account of what it is to do things for reasons, one that entails that some automatic processes themselves can cause us to do things for reasons,

independently of their relationship to our reflective capacities. What's more, the problems with the reflection views that I discussed in the last section suggest some further constraints on what such an account might look like, and more specifically I think they suggest that we should be after an account that grounds our ability to do things for reasons in psychological mechanism that we've acquired through some kind of learning process but that also recognizes that, under the right conditions, these mechanisms will cause us to do things for bad reasons, even by our own lights. As it turns out, recent research on reinforcement learning gives us a sense of what such mechanisms might be, and that's what I'll turn to now.

Chapter Three

Rationality Without Reflection

In the last chapter, I distinguished between what I take to be the two most common accounts of rational guidance. The first were reflection views, and the second were recognition view. In this chapter, I want to defend a version of the recognition view. In particular, I'll claim that (1) our ability to do things for reasons only requires us to recognize and respond to things as reasons and that (2) we often satisfy these conditions despite what research in social psychology suggests.

But defending these claims isn't an easy task. For one thing, it might be though that our ability to recognize things as reasons requires us to have access to those reasons, and so given the we often lack such access, it would seem to follow, on this view, that we often don't do things for reasons. But even if our ability to recognize reasons doesn't involve an access requirement, it might still be thought that, because of how arbitrary many of the stimuli in research in social psychology are, we often don't act in response to our recognition of reasons. Instead, we're just being pushed and pulled by arbitrary features of our environment, and there's no rational guidance to be found.[1]

However, it's both of these claims that I think we should resist. First, I don't think that we should assume that our ability to recognize things as reasons requires us to have access to the reasons that we recognize. Instead, it only requires a kind of *implicit* recognition. Second, I also think that we should resist the temptation to think that we often do things in response to arbitrary stimuli. In the first instance, this is because it's not clear that many of the stimuli in research in

[1] Both of these arguments can be found in Haidt (2001), for example.

social psychology are as arbitrary as they seem. But even if they are, a broader empirical perspective suggests that at least some automatic processes are much more intelligent and indeed reasons-responsive than this research suggests. And what I want to argue is not only that we often act on the basis of such mechanisms, but that when we do, there's a perfectly good sense in which we can be said to do doing things because we recognize a reason to do them.

1. Recognizing Reasons

Let's start with a brief overview of recognition views. In general, the idea here is that when we do something for a reason, we have to satisfy two condition: (1) we have to recognize something as a reason and (2) this recognition has to guide our behavior in the right sort of way. As I said in chapter 1, (2) raises interesting and difficult questions about the downstream or volitional aspects of rational guidance, but since what we're concerned with are the informational aspects, I'll focus on (1).

So, what is it to recognize something as a reason? Although different versions of the recognition view will have different answers to this question, the most common view is that it consists in some kind of normative belief or judgement. Thus, according to this view, when we recognize something as a reason, we believe or judge that it's a reason.[2] In other words, on this view, reasons enter into explanations of our behavior *via* our beliefs or judgments. So, for example, if my reason for not kicking you in the shin is that it will cause you pain, then on this version of the recognition view, that's because this consideration was the content of a belief that helped to guide my behavior.

[2] I take it that Scanlon (2000), Dancy (2000), and Raz (2002) hold a version of this view, just to name three prominent examples.

But despite their similarities, advocates of this view disagree about whether our ability to recognize reasons requires us to be aware of those reasons. For example, Elizabeth Anscombe (1953) says that we always have access to the reasons for which we do things in the sense that we would be able to report accurately on those reasons if someone were to ask us what they were. Likewise, Jonathan Dancy (2001) tells us that reasons-explanations of behavior require "that those features [i.e., our reasons] be present to the agent's consciousness – indeed, that they somehow be conceived as favoring the action" (p. 129), and Malissa Barry (2007) has more recently written that:

> When acting rationally, an agent undertakes to act in light of her belief about what she has reason to do. She chooses her action because it is supported by reasons. In this sense, rational action seems to embody a distinctly rational form of motivation in which the agent guides herself by the thought that an action is recommended by reason. This guiding thought need not always be explicitly articulated. For rational action to be possible, however, the agent must, at some level of awareness, conceptualize the features to which she is responding as reason-giving. (p. 232)

Although neither Dancy nor Berry clarify what they mean by "consciousness" or "awareness," I take it that they would accept something like Anscombe's view– that we can always report accurately on the reasons for which we do things if asked. I'll continue to call this the *access requirement on rational guidance.* In short, this is the claim that agents have to have access to the reasons that they recognize in the sense that they would be able to report accurately on those reasons if asked.

But not everyone who holds a version of the recognition views accepts the access requirement. For example, Fischer and Ravizza (1998; p.86) explicitly deny that we have to be aware of the reasons that we recognize, and T.M. Scanlon (2000; p. 23) seems to agree. So, I take it that the access requirement is optional. You can think that our ability to do things for

reasons requires us to recognize reasons without also thinking that we have to be aware of the reasons that we recognize.

Moreover, there seem to be good reasons to reject the access requirement. First, it might be thought that, when combined with research on automaticity, this requirement entails a kind of skepticism about rational guidance. After all, one of the main takeaways from this research seems to be that we often aren't aware of the causes of our attitudes and actions. So, if our ability to do things for reasons requires us to recognize things as reasons, and if we can't recognize something as a reason unless we're aware of it, it would seem to follow that we don't often do things for reasons.

In the next chapter, I'll argue that we might not be as ignorant of the causes of our attitudes and actions as research on automaticity might seem to suggest. So, I don't think that this is the most compelling argument against the access requirement. Still, there is an important kernel of truth in the argument above. Many times, if not typically, when we act on the basis of automatic processes, we won't be able to say what the causes of our attitudes and actions are. However, in at least a wide range of those cases, it'll also seem like we're doing things for reasons. So, even if the access require doesn't entail that we hardly ever do things for reasons, it'll still be under-inclusive.

One way to make this point is to consider various types of skilled activities. [3] For example, when an elite running jukes an on-coming defender or when an elite point guard makes a perfectly timed and placed pass, they often don't seem to know precisely what they're doing or

[3] See, for example, Michael Brownstein (2014).

why they're doing (at least not at the time of action).[4] Still, it's hard to believe that they're not doing these things for a reason. Indeed, elite athletes have spent endless hours training so that they can aptly respond to a wide range of situations without having to think too carefully about what they should do or why they should do it. So, it's plausible to think that, in many of these cases, the agent doesn't have access to the reasons for which they do things, and yet they're responding to reasons all the same.

However, I suspect that advocates of the access requirement might respond to this worry by pointing out that most of these are case of *intuitive* decision-making. That is, they're case in which the agent is responding to his or her sense that some action is called for or worth doing. What's more, they could claim that intuitions (or their contents) can be the reasons for which we do things. For example, if we were to ask an elite point guard why she made a certain pass, she might not be able to give a fully satisfying answer, but she could presumably say something like, "Well, it just seems like the right pass to make at the time." According to this response, that would be the reason for which she acted. So, on this view, as long as it's true that (1) elite athletes are often acting in response to their intuitions, (2) they typically (or even necessarily) have access to their intuitions,[5] and (3) their intuitions can be the reasons for which they do things, it follows that they will often have access to the reasons for which they do things. Thus, the access requirement is safe.

[4] Testimony from elite athletes seems to back this up (Brownstein, 2014). For example, here's Walter Payton, arguably the greatest running back in NFL history: "People ask me about this move or that move, but I don't know why I did something. I just did it." Similarly, Larry Bird, the legendary Boston Celtics player, once said that "[a lot of the] things I do on the court are just reactions to situations…A lot of times, I've passed the basketball and not realized I've passed it until a moment or so later."

[5] Intuitions are widely taken to be conscious mental states. See, e.g., Haidt (2001). So, on this definition, it's analytically true that we will be aware of our intuitions. I say "typically" above just to leave open the possibility of an alternative definition.

I have no issue with (1) and (2), and I'm willing to accept (3) if only for the sake of argument. However, I still don't think this response will save the access requirement, and that's because not all intuitions are created equal. Indeed, some of them are had for reasons, whereas others are. But if we accept the access requirement and assume, reasonably enough, I think, that we often lack access to the causes of our intuitions, then the access requirement won't be able to distinguish between cases where our intuitions are responsive to reasons and cases where they're not. So, in other words, while I'm willing to grant that we might always have access to the reasons for which we act, we won't always have access to the reasons for our attitudes, and in particular our intuitions.

Again, a good place to look to find examples of this is expert decision-making. After all, there's a world of difference between an elite point guard who senses that the only way that she can pull off a pass that she needs to make is by throwing it behind her back and the novice who has the same intuitions but for dubious reasons. The first is making her decision thanks to highly trained, experience-based processes, whereas the latter, for all we know, is just responding to a poster that he saw of someone throwing a behind-the-back pass. Still, it's plausible to think that, were we to ask both of these players about the causes of their intuitions, neither would be able to say. So, notice that, as long as you think that there is a significant difference in the rational status of these intuitions, which I certainly think there is, the access requirement won't be able to account for it.

One thing that this suggests, if I'm right, is that we shouldn't accept the access requirement when it comes specifically to the reasons for our intuitions. Thus, according to this view, there will be times when we recognize reasons to do things, and this recognition gives rise to an intuition that the relevant action is called for, but we won't be able to say what exactly the

reasons that we recognized were. When this happens, I'll say that we *implicitly* recognize those reasons. So, for example, on this view, we could say that the difference between the experienced point guard's intuitions and the novice's intuitions is that the experienced point guard's intuitions are had in response to her *implicit* recognition of reasons to throw a behind-the-back pass, whereas the novice's aren't.

Moreover, I think that this sort of view – that we can so something for a reason as long as we're acting in response to our implicit recognition of reasons – gives us a promising way to square research on automaticity with the idea that we often do things for reasons. Indeed, if most of what we think, feel, and do is based on processes that have more in common with the processes that support the point guard's decision-making than they do with the processes that support the novice's, then we might think that we often do things because we implicitly recognize reasons to do them.

However, there's an obvious problem with this suggestion, which is that research on automaticity (and in particular research on priming effects) makes it seem like the processes behind our attitudes and actions are much more like the processes behind the novice's decision-making than the expert's. After all, if we take studies at face value, then it looks like we often do things in response to arbitrary stimuli. That's not what the elite point guard is doing; it's what the novice is doing. So, before we can say that we often do things in response to our (perhaps only implicit) recognition of reasons, we need some reason to think that most of our attitudes and actions issue from processes that, while perhaps automatic, are responsive to significant features of our environment.

2. The Rationality Debate

But how might we show this? One way would be to argue that the stimuli used in many of these studies aren't as arbitrary as they seem. To be sure, the consensus in psychology seem to be that most of these stimuli are arbitrary. But a number of psychologists have argued that we should resist this claim. So, it's at least a matter of debate whether these studies show that we tend to be irrational.

To see an example of this debate, consider research on framing effects. As you might recall, what these studies show is that people's preferences can vary depending on how their options are framed. For example, people tend to prefer beef when it's described as 90% lean as opposed to 10% fat. Now, because these frames are logically equivalent, it's often thought that our preferences shouldn't be sensitive to which one is used. However, as Shlomi Sher and Craig McKenzie (2008) point out, even if two frames are logically equivalent, they can still convey different information. After all, there's a difference between saying a cup is half full versus saying that it's half empty; the fact that the speaker decided to use one of these descriptions over the other can often tell us something important about their attitudes. So, before we can conclude that framing effects are irrational, what we need to know is whether the frames are *informationally* equivalent as well.

If Sher and McKenzie are right, then there are two related questions that we might want to ask about research on priming affects. First, are the frames used in these studies informationally equivalent (or, alternatively, do they convey choice-relevant information)? Second, what happens to framing effects when informational equivalence is controlled for? Do they go away or remain the same?

Sher and McKenzie have focused almost all of their efforts on trying to answer this first question. Since there have been so many studies on framing effects over the years, they haven't been able to test whether all of the frames that have been used involve "information leakage," as they call it. But they have tested some of the most common types of frames, and their research suggests that they do convey choice-relevant information. In particular, what they found was that positive frames tend to serve as implicit recommendations, whereas negative frames serve as implicit criticisms. So, when the participants are asked to evaluate the options in question, it makes sense that their preferences would be congruent with the valence of the relevant frame, given that they know very little about the issue besides what the frames say and the pragmatic implications they generate.

Unfortunately, because there haven't been any studies that directly test what happens to framing effects when the pragmatic implications of the frames are canceled, the second question is harder to answer. However, in a recent review of their research, Sher and McKenzie (2008) point to studies in other domains that shed some light on this issue. For example, research on hypothesis testing shows that people tend to take observations that are explicitly mentioned by a hypothesis to provide more support for it than do observations that aren't explicitly mentioned, even when the evidential value of the two observations is the same. This effect is often seen as evidence of confirmation bias, but McKenzie and Mikkelsen (2000) argue that it's instead the result of a framing effect. More specifically, they found that people tend to assume that hypotheses are usually stated in terms of rare events, and when this assumption is combined with the fact that, from a Bayesian perspective, rare events that confirm a theory can provide more support for it than do confirming common events, the participants' behavior turns out to be surprisingly rational. However, once it's made clear to them that they shouldn't make this

assumption, their tendency to prefer the stated observation is less pronounced. This isn't a classic example of a framing effect, but it does suggest that they too might decrease once the pragmatic implications of the frames are canceled.

If Sher and McKenzie are right, then I think we can draw a valuable lesson about the rationality of framing effects. At first glance, the stimuli in these studies really do seem arbitrary, and so it's easy to conclude that the participants' responses are irrational. However, on closer inspection, this doesn't seem to be the case. In fact, it appears that these effects involve a good deal of information processing. The participants are responding not only to the frames that they've been given, but also to the fact that the researchers decided to use those frames instead of other frames, which, as Sher and McKenzie stress, can be relevant information. So, before we say that the stimuli in these studies are arbitrary, we need to make sure that we know what the relevant stimuli are. In this case, it looks like the participants are tracking more information than people often think.

Although I find Sher and McKenzie's research on framing effects especially convincing, it's important to note that it's far from the only research in this vein. Indeed, there are many other examples of effects in psychology that are arguably more rational than they might at first seem. I already mentioned one of them with McKenzie and Mikkelsen's work on hypothesis testing (but also see Klayman & Ha (1987) and Oakford & Chater (1994) for similar arguments), and other examples include Gigerenzer's (1991) arguments against base-rate neglect, conjunction fallacy, and overconfidence bias, and Payne, Bettman, and Johnson's (1993) finding that choice-strategy behavior (i.e., our tendency to stop evaluating our options carefully once we have more than two or three of them) is actually a reasonable trade-off between accuracy and effort (Beach & Mitchell (1987) make a similar point). In all of these cases, the upshot is very much the same:

what's often said to be irrational behavior turns out to be quite rational as long as it's evaluated using the right normative model.

We can also find examples of this in research on moral cognition and behavior. Take, for instance, Haidt's research on moral dumbfounding. In these studies, the participants are given vignettes that have been carefully written so that the actions that they describe don't obviously violate any moral norms. For example, in the much-discussed incest vignette, they're told that Mark and Julie, the brother and sister, agree that "it would be interesting and fun if they tried making love," use multiple forms of birth control, decide not to do it again, and promise to "keep that night a special secret." So, when the participants fail to say why they think that Mark and Julie's behavior was morally wrong, Haidt takes this to show that their judgments aren't based on rational processes but are instead driven entirely by a reflex-like disgust response to incest. However, as Peter Railton (2014) notes, what Mark and Julie did was still risky and poorly motivated (after all, they decided to jeopardize their relationship and future simply because they thought it would be "interesting and fun" to sleep with each other), and both of these features of the case are certainly relevant to our assessment of their decision, even if things ultimately did turn out okay for them. Consequently, it's plausible to think that, although the participants aren't able to say what they are, their judgments that Mark and Julie's behavior is wrong could still be based on reasons.

A different but no less important lesson can be draw about studies on incidental disgust. Recall that these are the studies that shows that disgust-eliciting stimuli can amplify our moral judgments even when they're irrelevant (or "incidental") to the judgment at hand. For example, sitting in a dirty desk has been shown to make people's moral judgments harsher (Schnall et al., 2008), and the same goes for smelling fart spray (ref). Since it's hard to see how these stimuli

could be anything but arbitrary, it's unlikely that we'll be able to give a rational explanation of these effects. Still, there's reason to question their size. Indeed, in a recent meta-analysis, Landry and Goodwin (2015) found that when they computed the average effect size of all the studies on incidental disgust that were reported on online databases (31 of which were published and 20 unpublished), it was still significant but quite small (d=.17), and they further point out that once publication biases are factored in as well, the effect all but disappears. So, even if there isn't a rational explanation for the effect of incidental disgust on moral judgments, it's probably not that large.

In sum, it's not clear that research in social psychology shows that arbitrary stimuli have a pervasive influence on our attitudes and actions. Of course, that's not to say that we're always rational; surely, we sometimes do things for bad reasons or for no reason at all. But no one has to deny that. Instead, the issue is here whether our attitudes and actions often have arbitrary causes, and when we take a closer look at the research in question, I don't think we find enough evidence to say that they do. In some cases, that's because what might seem like arbitrary stimuli turn out to be quite meaningful. But in other cases, it's because the sizes of the relevant effects might not be as large as they were first reported to be. In any event, it's a big debate among psychologists whether studies like the ones discussed above show that humans tend to be less rational than we would like to think, and I don't think we should take it for granted that they do. Maybe, in the end, we should accept this claim, but it's got to be as the conclusion of an argument and not just a premise.

3. Skilled Action and Model-Based Learning

Still, even if we take these studies at face value, there are other reasons to think that we often do things on the basis of psychological processes that, while automatic, are tracking

meaningful features of our environment. The argument here comes from Peter Railton (2009; 2014),[6] and as I understand it, it involves two key observations: first, that most of our everyday exercises of agency draw on a whole suite of social, emotional, and practical skills, and second, that recent research on reinforcement learning suggests that these skills are subserved not by the kind of crude, reflex-like processes that have long been associated with automaticity but by "model-based" learning systems that guide our behavior through complex representations of our environment that get continuously updated through experience. And if both of these claims are right, then I think we should say that we're often much more like the elite point guard than the novice – that is, we're often acting on the basis of processes that allow us to respond quickly but flexibly to meaningful changes in our environment.

Let's start with the idea that most of our everyday exercises of agency draw on a wide array of social, emotional, and practical skills. Railton (2009) isn't the first philosopher to make this observation, of course,[7] but his discussion contains a distinction that I find useful. First, he says that exercises of agency often involve what he calls "agent competence." By this, what he has in mind is:

> [A] set of skills akin to the skills that equip someone to be an effective administrator, *inter alia*: an ability to focus attention selectively, but also to be somewhat mindful of several things at once; an ability to set goals and make plans, but also to revisit and revise them in light of experience; an ability to interpret situations and form a view about what's at stake and which actions are in the offing; an ability vividly to imagine alternatives and identify necessary means toward them as well as their remoter consequences; an ability to deliberate and decide among options; an ability to adjust means to ends; an ability to maintain morale and generate motivation in the face of difficulty; and so on. (p. 86)

[6] For similar arguments, though, see Patricia Churchland and Chris Suhler (2014), and John Allman and Jim Woodward (2007).

[7] Aristotle, for example, famously thought of virtue as a practical skill (see Annas (2011) for an interpretation of his view). For a different perspective on the role that skill plays in human agency, see also Dreyfus (2005).

As Railton goes on to note, all of these skills play an integral role in practical rationality. For example, in order to be even a minimally decent practical agent, you have to know when to deliberate and when to stop, and you have to know which considerations are relevant to your deliberation and which ones aren't. However, according to Railton, these skills aren't usually taught explicitly. Instead, we develop a mastery of them though implicit learning processes, just like we often become competent language users through implicit learning processes as well (pp. 84-85).

Second, Railton points out that our everyday exercises of agency involve what he calls "practical intelligence," which refers to "the ability to solve a range of problems specific to the exercise of agency" (p. 89). For example, when we're interacting with other people, we often have to quickly read their minds and anticipate how they'll respond to our behavior. We also have to have some (perhaps only inchoate) sense of the norms that are governing our interaction and what they demand of us in that situation. As Railton argues, such "people skills" are an indispensable part of competent agency, and yet they too are often developed and deployed outside of our awareness.

Taken together, what both of these claims suggest is that in order to be competent practical agents, we have to know, among many other things, how to use our reflective and deliberative capacities wisely and how to interact with other people appropriately. What's more, as Railton argues, we typically exercise these skills in ways that are quite selective and sensitive to even the subtlest changes in our environment. For example, we all have a pretty reliable sense of when our friends and family members could use our support and when they should be left alone, but in most cases, we have access to only a fraction of the information that causes us to make these decisions. So, if Railton is right, then it seems like the kinds of mental systems or

processes that support our everyday exercises of agency have to be quite flexible and capable of responding quickly but intelligently to subtle changes in our environment. But what kind of systems or processes might these be?

This is where recent develops in reinforcement learning can help us out. In particular, over the last decade or so, it has become increasingly clear that most of our decisions reflect the operations of two different kinds of reinforcement learning systems, a computationally cheap but inflexible "model-free" system and a computationally more expensive but also more flexible "model-based" system.[8] As Railton argues, only model-based systems would be able support the kind of decision-making that's characteristic of experts, and since our own decision-making typically issues from a broad range of skills, it's plausible to think that it too will be largely supported by model-based systems.

To get a better sense of how these two systems work, I think it helps to consider a simple reinforcement learning problem, seen in figure 1. For lack of a better name, let's call the agent here, represented by the image of the face, "Smiley." Smiley's task is simple: it has to learn how to get from the start of the maze to the reward state, represented by the box with the star, in the fewest moves possible (i.e., by moving all the way down to state 22 and then taking a right). How can it do this?

[8] See Crockett (2013) and Cushman (2013) for recent discussions of these developments. Interestingly, they use the distinction between model-based and model-free systems to independently develop very similar theories of moral cognition.

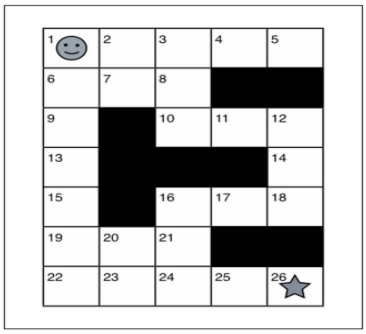

Figure 1: Simple Reinforcement Learning Problem

from Cushman (2013)

There are two different approaches that we might take. On the model-based approach, we would give Smiley the ability to represent not only where it is in the maze and the moves that are immediately available to it from that state, but also the consequences of those moves and the value of their outcomes. That way, when Smiley is trying to figure out where to go, all it would need to do is consult these representations (also known as a "causal model") and select the course of action that it thinks will lead to the reward state in the fewest possible moves. However, Smiley obviously won't have all of this information available to it at first, and so it'll have to learn what the layout of the maze is. On the model-based approach, the way that we do this is by giving Smiley the ability to update its causal model in light of any discrepancies between Smiley's experiences and its representations of the maze. As a result, once Smiley has explored the maze enough times, it should develop an accurate causal model, which it will then be able to use to navigate the maze efficiently.

On the model-free approach, by contrast, we don't give Smiley nearly as many representational resources. Instead, we only give it the ability to represent where it is in the maze, what actions are immediately available to it in that state, and what the value of those

actions are. However, because Smiley can't represent the outcome of those actions, it doesn't know why the actions the value assignment that they have. All it knows, from any given state of the maze, is that one of the moves available to it is better than the other, and its decisions will reflect these value assignments.

Still, even though model-free systems aren't able to represent the outcomes of their actions or the value of those outcomes, this version of Smiley would still be able to navigate the maze efficiently if we could also design it to learn how to build the value of an action's outcome into its evaluation of the action itself. On the model-free approach, the way that we do this is by using what's known as "temporal distance reinforcement learning," which works by getting the agent to treat predictors of rewards (and predictors of predictors of rewards, etc.) as though they're rewards themselves. Thus, once the model-free Smiley reaches the reward state, it'll learn to treat being in the state that immediately preceded it (in this case, state 25) as rewarding itself, and then it'll learn to treat being in the state that immediately preceded that state (i.e. state 24) as rewarding as well, and so on, until it is able to string together a series of value assignments that corresponds to the most efficient route. So, for example, once Smiley has explored the maze enough, it'll learn that moving from 1 to 6 is better than moving from 1 to 2, and that moving from 6 to 9 is better than moving from 6 to 7, and that moving from 9 to 13 is better than moving back to 6, etc., although it won't know why they have these values.

Since model-free systems don't require the agent to represent the outcomes of their actions and the value of those outcomes in order to figure out what to do, they're computationally much cheaper to use than model-based systems are. But at the same time, because of their lack of representational resources, they're a lot less flexible. Thus, if we were to move the reward state from 26 to (say) 18, it would take the model-free version of Smiley a lot

longer than the model-based version to learn how to navigate the maze efficiently. After all, in order to make the relevant adjustments, all the model-based system would need to do is update its map of the maze to reflect the new location of the reward. However, the model-free system would need to recalibrate all of the values that it assigns to actions by working its way backwards again from 18 to 1. As a result, even after the model-free system has encountered the new reward state, there will still be times early on in this process where it expects moving from 1 to 6 to be more rewarding than moving from 1 to 2.

The distinction between these two systems was first introduced by research on machine learning, but it's now pretty clear that they're not unique to robots. In fact, it appears that most foraging animals, including humans, have learning systems that operate in very similar ways (Doll et al., 2013). For example, studies involving both humans and other kinds of animals have shown that dopaminergic neurons in our midbrains fire in patterns that are consistent with temporal distance reinforcement learning – they first respond only to rewards, but over time they start to respond to predictors of rewards and then to predictors of predictors of rewards, and so on, suggesting that we learn to treat those predictors as being rewards themselves (this is why, for example, the smell of coffee or the sight of a Starbucks logo can itself be a source of pleasure) (Fiorillo, Tobler, & Schultz, 2003; Schultz et al., 1997; Schultz & Dickinson, 2000; for studies involving humans, see McClure, Berns, & Montague, 2003). Similarly, many studies have shown that rats are able to navigate mazes with remarkable efficiency by relying on internal maps of their environment that get continuously updated through experience (Doll *et al.,* 2013). Interestingly, it also appears that rats will explore these models in their sleep, which in turn allows them to take shortcuts in mazes that they otherwise wouldn't have known about (Ji &

Wilson, 2007; Gupta et al., 2010).[9] Finally, and perhaps most importantly, studies have also shown that model-based learning plays a key role in expert decision-making. For example, it appears that what separates elite athletes from very good ones isn't that they have superior physical traits or more practiced motor skills, but that their internal models are more accurate, complex, and available for them to draw upon when it counts (Yarrow, Brown, & Krakauer, 2009).

Of course, talking about model-free and model-based systems as though they're completely independent of each other can be misleading. Indeed, there's good reason to think that they can and often do interact and that many of our decisions are a product of their joint operations (Cushman, 2013). Still, they can come apart, and that's why they're often regarded as distinct systems. For example, studies on "reward devaluation" have shown that once rats have been overtrained they will continue to press levers that they associate with food even when they're not hungry anymore. But if they haven't been overtrained, their choices will be sensitive to changes in their motivational states (e.g., hunger levels). According to the standard interpretation, what these results suggest is that the overtrained rats are being guided by model-free systems, whereas the other rats are being guided by the more flexible model-based systems (Balleine & Dickinson, 1998).

So, what's the relationship between these two systems and our everyday decision-making? Because model-based decision-making resembles what we would ordinarily think of as a paradigm form of practical reasoning (after all, these systems work by selecting actions based

[9] In the middle of last century, Edward Tolman (1948) noticed this as well, and it led him to argue that, contrary to behaviorism, classic conditioning alone can't teach rats how to navigate mazes; instead, he posited that they also have to develop and employ internal maps of their environment, or else we can't explain how they take novel routes to rewards.

on their expected utility), people often assume that it's conscious.[10] As a result, they also often assume that, because our everyday decision-making is largely intuitive, it must be the result of model-free systems. And when you combine this assumption with all of the research that seems to suggest that the processes behind our attitudes and actions are relatively dumb, it starts to look like a pretty safe bet.

However, Railton (2016) has recently argued that this is a mistake. Instead, on his view, most of our intuitive decision-making is the result of *implicit* model-based systems. In part, this is because it seems like it's a mistake to assume that model-based decision-making has to take place consciously. As we've already seen, there's ample evidence that many other kinds of foraging animals rely on this kind of decision-making to navigate their environment, and while these animal might be more intelligent that we often assume, it's hard to imagine that they are explicitly thinking about their options and weighing their long-term effects. Instead, it's much more plausible to think that their decisions are being guided by implicit model-based systems. .

What's more, Railton also points out that it's hard to see how model-free systems alone would be able to account for the kind of flexible-yet-spontaneous decision-making that we often carry out in our everyday lives. Indeed, we've seen that there's evidence to support the idea that skilled behavior in general depends on model-based learning, and if Railton is right that many of our everyday exercise of agency depends on a broad range of practical and social skills, that's pretty good reason to think that these systems will play an integral role in our everyday decision-making. However, because that decision-making often takes place outside of our conscious

[10] See Brownstein (2016) and Greene (2016). Interestingly, because Brownstein assumes that model-based decision-making must be conscious, he reads Railton as defending the claim that model-free systems are more sophisticated than people often assume and are capable of supporting skilled intuitive decision-making. However, Railton (2016; and in person conversations) makes it clear that his view is that intuitions are summary outputs of implicit model-based processes.

awareness, Railton concludes that it must be based, at least in large part, on implicit model-based systems.

If Railton is right, then I think what we end up with is an account of practical rationality that has clear parallels with certain views about epistemic justification. In particular, it seems like the practical analog of views that say that our beliefs are justified insofar as they issue from processes that track the truth. Likewise, Railton seems to think that we can be said to do something for a reason insofar as our actions issue from processes that are sensitive to reasons (and as it turns out (i) model-based systems are sensitive to reasons and (ii) they guide most of our everyday decisions).

But I want to go one step further than this and try to fit his view within the framework of recognition views. This is partly because I think it'll be difficult to give a fully adequate account of what it is for a process to be "sensitive" to reasons. But if it can be shown that when we act on the basis of model-based processes, we are usually doing things because we recognize that we have reason to do them, very few people will balk at the suggestion that these processes allow us to do things for reasons. However, I also want an account of rational guidance that's applicable to traditional concerns in ethics, and those concerns are almost always understood in familiar folk psychological terms. So, the other reason why I want to situate Railton's view within a recognition framework is that it will be easier to apply to debates about moral responsibility and other issues in ethics.[11]

[11] Consider, for example, Fischer and Ravizza (1998). On their view, we're responsible for our actions just in case they issue from reasons-responsive psychological mechanisms, where they analyze reasons-responsiveness in terms of our ability to recognize and respond to reasons in certain counterfactual situations. However, since they don't elaborate on what it is to recognize something as a reason, it's not obvious what the relationship between their view and Railton's view might be. What I take myself to be doing here is bridging this gap by showing that when we act on model-based systems we can normally be said to recognize something as a reason in the sense that Fischer and Ravizza have in mind.

4. Recognizing Reasons Implicitly

Thus, what I want to show in the rest of this chapter is that when we act on the basis of model-based systems, we're usually acting in response to our recognition of reasons. What's more, since I assume that most people who hold a recognition view would agree that we can recognize something as a reason as long as we believe that it's a reason for us to do that thing, the claim that I want to defend more specifically is that when we act on the basis of model-based systems, we can be said to do those things because we *implicitly believe* that we have reasons to do them.

However, before I try to defend this claim, it might help to consider a few examples of the sort of cases in which I think it's plausible to say both that our behavior is being guided by model-based systems and that we are doing things because we implicitly believe that we have reasons to do them.

The first case comes from Susanna Siegel (2016). Here's how she describes it:

> [**Kindness**] The person ahead of you in line at the Post Office is finding out from the clerk about the costs of sending a package. Their exchange of information is interspersed with comments about recent changes in the postal service and the most popular stamps. As you listen you are stuck with the thought that the clerk is kind. You could not identify what it is about the clerk that leads you to this thought. Nor could you identify any generalizations that link these cues to kindness. Though you don't know it, you are responding to the combination of what she says to the customer, her forthright and friendly manner, her facial expressions, her tone of voice, and the way she handles the packages. (p. 95)

Although Seigel introduces this case to argue that we can make inferences without being explicitly aware of the information on which they're based, I think it just as nicely illustrates the kind of processes that are responsible for most of our social cognition. After all, what you're doing here is incorporating vast amounts of information about someone in order to arrive at an

overall judgment of their kindness, and yet you're barely aware of that information or how it factors into your assessment. Still, the features of the clerk to which you're responding are highly relevant to ascriptions of kindness, and it seems like there must be some sense in which you take them to count in favor of thinking that she is kind. In this case, then, I would say that you judge that the clerk is kind because you implicitly believe that the features that your judgment is tracking are reasons for this belief.

The second case comes from Michael Brownstein (2016). It's similar to Siegel's case, but it comes from real life and, as Brownstein himself says, nicely captures the kind of social expertise or "fluency" that Railton thinks is characteristic of competent practical agency. Here's how he puts it:

> [**Obama**] In front of millions of viewers, Obama and Chief Justice of the Supreme Court John Roberts both fumbled the lines of the Oath, opening the possibility for a disastrously awkward moment. But after hesitating for a moment, Obama smiled widely and nodded slightly to Roberts, as if to say, "It's okay, go on." These gestures received little explicit attention, but they defused the awkwardness of the moment, enabling the ceremony to go on in a positive atmosphere. Despite his nervousness and mistakes, Obama's social fluency was on display. Most of us know people with similar skills, which require real-time, fluid spontaneity, and can lead to socially valuable ends. (p. 301)

Granted, we don't know exactly what was going on in Obama's head when he decided to smile and nod at Roberts, but we can easily imagine some of the factors that might've caused him to do this: everyone's expectation that, as the new President, he's the one who needs to diffuse the awkwardness of the moment; the pained look of humiliation on Roberts' face; the sense that, if he just plays it cool and doesn't make a big deal of Roberts' mistake, it wouldn't receive much attention in the news that day and would save his Chief Justice from having to suffer even more embarrassment; and so on. And we can further imagine that, just like you in the example above,

Obama wouldn't be able to cite these cues as the reasons for his gesture. But again, it seems that, at some level and in some way, he must represent them as calling for some kind of response, and to capture this idea I would say that Obama believed that he should smile and nod at Roberts because he implicitly believed that the fact that it was his responsibility to do something, that Roberts looked embarrassed, and that he didn't want Roberts to endure more embarrassment were reasons to do this.

The last case comes from Burns, Monteith and Parker (2017). Unlike the first two examples, it's not a success story. It's a case where intuitive decision-making leads the agent astray:

> [**Implicit Bias**] Andy, a White man, is walking through a crowded mall. Like many people, Andy thinks of himself as a tolerant, fair, and egalitarian person. As Andy is walking, he happens to see a Black man holding a purse. The thought, "I bet he stole that purse" automatically enters Andy's mind. Just then, he sees a woman emerge from a store, take the purse, and put it over her shoulder. Suddenly, Andy realizes that he stereotyped the Black man as a criminal. (P. 99)

It might be surprising to find an example of implicit bias here, particularly since they've long been thought of as the products of associative learning mechanisms and not the sort of flexible, model-based processes that I take to ground our ability to do things for reasons. But recent research has shown that implicit biases can be modulated in ways that we wouldn't predict if they were based entirely on associative mechanism (Mandelbaum, 2016), and this has led some commentators to speculate that model-based processes must be implicated as well (Huebner, 2016; Brownstein, 2016).[12] As a result, I'm inclined to think that we can give a by-now familiar kind of belief-based explanation of Andy's behavior in this case. So, why did he think to himself that the Black man probably stole the purse? As I see it, it's because he implicitly believes that

[12] I'll elaborate on this view about the nature of implicit biases and the evidence that supports it some more in the next section.

the fact that men don't usually carry purses and that Black men are especially likely to commit crimes are reasons to have this belief.

5. Beliefs and implicit biases

But why do I think that we should make these belief ascriptions? One of the reasons is that the sort of representations that model-based systems use to navigate the environment have some important features in common with beliefs. For example, beliefs are widely thought to be (perhaps *inter alia*) (i) representational states that are both (ii) action-guiding and (iii) evidence-sensitive, and that seems to me to be pretty good description of the kinds of information states that are used by model-based systems. After all, these states can flexibly guide our behavior by representing the actions available to us and the value of their outcomes in a way that's apt to be revised in light of any discrepancies between the way that they represent our environment and the way that it really is. So, if these information states walk and talk like beliefs, why not think that they actually *are* beliefs?

Another reason to think that they are beliefs is that some prominent views of beliefs seem to entail that they are. A good example of this is the sort of representationalist view that Dretske (1988) defends. On his view, and to put it crudely, someone can be said to believe that P when they have mental systems whose function is to indicate, thanks in part through learning processes, whether or not P is the case. Although the kinds of learning processes that Dretske has in mind are associative (he was writing well before research on model-free and model-based learning had taken off), there's no reason to think that they have to be. So, on his view, it would seem to follow that the kinds of information states that model-based systems use to guide our actions would count as beliefs as well.

The same could also be said about certain dispositional accounts of beliefs. On these views, to believe that P is just to have a certain set of dispositions. Exactly what those dispositions are is a substantive question, but on more "liberal" views they include not only observable behavioral tendencies (e.g., the tendency to look in the fridge for a drink when you're thirst or to say "Hey, it's a harvest moon!" when you see a big, orange orb in the sky), but also tendencies that aren't as easy to observe (e.g., the tendency to feel certain emotions in response to certain stimuli or to think to yourself "I don't want to get out of bed" when you don't want to get out of bed).[13] Now, since these views identify beliefs with sets of dispositions and not the psychological structures that ground them, they wouldn't say that the representations that model-based systems use to guide our behavior are beliefs. But I take it that they would (or at least could) say that many of the behavioral tendencies, broadly understood, that these systems support are enough to warrant the relevant belief ascriptions. So, for example, if Andy is inclined to think to himself "I bet he stole the purse" when he sees a Black man holding a purse, and if he's inclined to cross the street when he's headed toward a Black man late at night, then we might think he really believes that Black men are prone to crime. Of course, given how Monteith sets up the case, it's safe to assume that Andy also has behavioral tendencies that militate against such belief ascriptions (e.g., a tendency to avow egalitarian values when asked about his political or moral views), and this certainly complicates matters. But the dispositions in which beliefs consist on these views aren't (or at any rate needn't be) all-or-nothing. It could be that you need to have only some but not all of the relevant dispositions to have the relevant belief. In addition, of course, it could be that Andy has inconsistent beliefs. So, I take it that there will at least be some versions of the dispositional view that would allow us to say that, despite his protests to the contrary, Andy really does have these racist beliefs.

[13] For this kind of dispositional account of belief, see Schwitzgebel (2002).

So, if we accept either of these views, then it seems like we'll be able to make the belief ascriptions that I want to make. That is, we'll be able to say that, for example, in the Kindness case, you genuinely believe that the clerk's tone and gentle manner are reasons to think that she's kind or that, in the Implicit Bias case, Andy genuinely believes that Black men are especially likely to commit crimes. But there are many other accounts of beliefs, and I don't want my defense of this position to rest entirely on disputed views about the nature of beliefs. Instead, I'd like to take a more ecumenical approach. So, rather than playing offense and trying to show that these states count as beliefs, I want to play a little defense and ask why we should think that they *don't* count as beliefs.

One person who thinks that these states shouldn't count as beliefs is Tamar Gendler (2008a; 2008b; 2011). This is especially true of implicit biases, which she explicitly says aren't beliefs. Instead, on her view, what's going on in the Implicit Bias case (and in the other two cases as well, presumably) is that Andy's behavior is being guided by what she calls an "alief." As she defines them, aliefs are complex mental states with affective, representational, and motivational parts. At once, they represent the way the world is, involve certain feelings, and mobilize action. For example, if you're watching a scary movie, then at some point you'll probably have an experience whose content is something like "Scary guy. Fear! Run away!" On Gendler's view, what you're experiencing here is an alief. It's not three different mental states working together to produce a coordinated response, but rather a single mental state with three different components – one representational, one affective, and one motivational. So, to return to the Implicit Bias case, Gendler would presumably say that it's not that Andy *believes* that Black men are likely to commit crimes. It's that he has an alief whose content is something like "Black man. Danger! Call the cops!" and that's why he has the relevant thought.

But what's wrong with the belief-based explanation? Gendler's main argument against such explanations turns on a prominent account of what beliefs are, which she attributes to David Velleman (2000). On this view, to believe that P is, roughly speaking, to have a mental state that not only represents P as being true, but is also sensitive to considerations that bear on whether P is, in fact, true. It's this last part about beliefs, which we might call their "evidence-sensitivity," that, according to Velleman, separates them from other kinds of representational states, such as assumptions or imaginings. After all, if you assume that P, then you don't have to be inclined to change your mind in light of countervailing evidence. The same goes for imaginings: if you imagine that P, then it doesn't really matter, from the perspective of this attitude, whether or not P is the case. But beliefs, according to Velleman (and Gendler, by extension), are different. As they see it, if you believe that P, then you'll at least be inclined to revise or abandon this belief in light of evidence to the contrary. So, on this view, whereas other kinds of representational states aren't necessarily regulated by considerations that bear on whether their contents are true, beliefs are. That's why, as G.E. Moore observed long ago, it sounds so odd to say "I believe that, but P is not the case," but it's not odd to say "I imagine that P, but P is not the case" or "I assume that P, but P is not the case."

With that said, let's return to the Implicit Bias case. From what I gather, there are two reasons why Gendler wouldn't say that Andy believes that Black men are especially prone to commit crimes (and thus takes this to be a reason to believe that the Black man he saw at the mall stole the purse). First, on her view, the attitudes that guides this thought don't have the right kind of content. Like most psychologists, Gendler is operating on the assumption that implicit biases are caused by and stored in associative mechanisms and structures. In general, what this means is that we not only acquire these biases by being repeatedly exposed to stimuli that pair

members of certain social groups (e.g., Black men) with stereotypical properties (e.g., *being dangerous*), but also that their structure is such that when one item in the associative network gets activated, the other items are likely to be activated as well, thereby priming us to think and act in stereotype-congruent ways. As a result, on this view, implicit biases don't have propositional content. Instead, their content is specified by the activated items in the associative network. Thus, according to this view, Andy's racist thought is caused not by the belief that Black men are dangerous, but rather by an attitude whose content is "Black man! Dangerous!," just as Gendler's view predicts. And if this is right, then just by dint of having the wrong kind of content, implicit biases wouldn't count as beliefs.

Second, Gendler also argues that implicit biases don't have the kind of evidence-sensitivity that beliefs have. In part, that's because they're often poorly integrated with our other attitudes. By now, it's a well-known fact that you can be explicitly committed to egalitarian ideals and still harbor implicit biases against people of many different races, genders, sexual orientations, sizes, and so on. So, implicit biases seem to be compartmentalized in a way that's not characteristic of beliefs. What's more, as Gendler notes, they also aren't very responsive to familiar mechanisms of belief formation and revision. Andy can rehearse to himself all he wants the reasons why he shouldn't believe that Black men are dangerous, and he'll probably still continue to have implicit biases against them. In fact, if the associative picture of implicit biases is right, then the way to get rid of implicit biases isn't to give someone a sound argument for thinking that they're wrong; rather, it's to modify certain contingencies in their environment (e.g., by repeatedly exposing them to stimuli that involve one but not both of the relevant relata). So, the other reason why Gendler denies that implicit biases are belief is that they're not responsive to the same methods of revision. As she sees it, to change a belief, you just have to

give someone reasons to think otherwise, but to change an implicit bias, you can't use the same rational tricks – you have to change their environment.

However, it seems to me that both of these arguments face serious challenges. First, as Eric Mandelbaum (2016) has recently argued, the view that implicit biases are stored in an associative structure is difficult to reconcile with the available evidence. As I already mentioned, one nice feature of the associative view is that it makes clear predictions about the conditions under which implicit biases can be extinguished. It says that the only way to get rid of these biases, aside from cutting out huge swaths of the agent's brain or taking other extreme measures, is to expose agents to different stimuli (that way, they'll learn not to associate members of the relevant group with the stereotypical properties). However, studies have shown that this isn't the case. Implicit biases can be altered by other (indeed quite rational) means. For example, Brinol and his colleagues (2009) showed that having people read good arguments for affirmative action policies can significantly improve their performance on race-based Implicit Association Tests, and an earlier study by Sechrist and Stangor (2001) found that implicit biases can be modulated by beliefs about what the majority of our peers think about the bias in question. As Mandelbaum argues, in both of these cases, it looks like the participants' implicit biases are behaving much more like beliefs than associative structures, which suggests that the purely associative story about the nature of these attitudes is probably false. Thus, insofar as Gendler's first argument for thinking that implicit biases aren't beliefs (because they lack propositional content) presupposes such a view, it's not very convincing.

Second, I also think we should question her argument from the evidence-insensitivity of implicit biases. For one thing, both of the studies above suggest that implicit biases are, at least to some extent, responsive to our beliefs about what we ought to believe. So, I think it's fair to

ask whether they are as evidence-insensitive as people, including Gendler, often seem to assume. For another thing, we might also doubt whether beliefs really are as sensitive to evidence as Gendler and others suggest. For example, there are lots of clinical cases where it seems like someone believes that P, and yet they're not at all inclined to change their mind in light of evidence that not-P. Just think about someone with Capgras syndrome: they might genuinely believe that their spouse is a robot, but you're not going to change their mind by giving them reasons to think otherwise. Likewise, people with Schizophrenia might know, on the one hand, not to trust the voices in their heads but, on the other, trust them all the same. So, it's fair to ask whether beliefs are evidence-sensitive in precisely the way that Velleman and Gendler assume. In both of these cases, it seems plausible to think that the attitudes in question really are beliefs, but it doesn't look like they're apt to be revised in light of evidence to the contrary.

What's more, we don't even need to consider clinical cases to make this point. As Andy Egan (2011) points out in response to Gendler, many of our own beliefs are neither fully integrated with our other attitudes nor ideally responsive to the mechanisms of belief formation and revision, and yet they're still beliefs. For example, you might know that, like most other professors, you tend to overrate your teaching but still continue to believe that you're an above average teacher nevertheless (Egan, 2005). Or I might sincerely believe that I have a winning lottery ticket or that I'll be eaten by a shark when I go swimming in the ocean, even though I know fully well that the odds of either happening are slim, to say the least. In any case, as many other people have already pointed out (for example, see Egan 2008; Gilbert 1991; Gilbert *et al.*, 1993; Huebner 2009; and Mandelbaum 2011), it seems like we all have beliefs that are compartmentalized, formed automatically, and incredibly recalcitrant. Thus, it can't be essential

to beliefs that we're always inclined to be abandoned or revised them in light of countervailing evidence. That's just too strict of a standard.

Just to be clear, the point here isn't that we should reject the idea that beliefs are, in some important sense, evidence-sensitive. Nor do I mean to suggest that implicit biases are as sensitive to evidence as paradigm examples of beliefs are. But what I do want to stress is just how difficult it is to distinguish between beliefs and implicit biases in terms of their sensitivity to evidence. If we make the standards of evidence-sensitivity so strict that it clearly rules out implicit biases, we'll also likely rule out many beliefs. But if we relax the standards so that all of our beliefs are included, as fragmented, automatic, and recalcitrant as they might be, then implicit biases are bound to be included as well. So, it seems to me that Gendler can't have it both ways. She can't have a plausible account of what beliefs are *and* claim that implicit biases aren't among them. Instead, it's one or the other – either she has a plausible account of what beliefs are and includes implicit biases in that category, or she can exclude implicit biases but be stuck with an implausible account of beliefs.

Like Gendler, Neil Levy (2015) also argues that implicit biases aren't beliefs. In some respects, his argument is an improvement on Gendler's. Importantly, citing Mandelbaum, he says that we shouldn't assume that implicit biases have an associative structure and so can't argue, just on the basis of differences at the level of their respective contents, that implicit biases aren't beliefs. But in other respects, his argument is quite similar to Gendler's. In the end, he still wants to say that implicit biases aren't beliefs because "their sensitivity and responsiveness to other mental representations is too patchy and fragmented for them to properly be considered beliefs" (p. 800). And at the heart of this argument are two claims about the nature of beliefs. First, he says that beliefs are "inferentially promiscuous inasmuch as the belief that p can interact

(appropriately) with any other propositional attitudes" (p. 805). Second, he tells us that beliefs are responsive to evidence, although he never says what exactly this means. At any rate, Levy sees both of these characteristics of beliefs as "two sides of the same coin," writing that "beliefs are inferentially promiscuous, causing the update of other beliefs, because beliefs are responsive to evidence" (p. 805). And since, on Levy's view, implicit biases aren't responsive to evidence, he concludes that they're not beliefs.

But because he still wants to claim that implicit biases and beliefs can be distinguished in terms of their sensitivity to evidence, Levy's argument is vulnerable to many of the same worries that Gendler's faces. To his credit, Levy admits that "the beliefs of actual human beings are far from perfectly responsive to evidence or apt to feature in inference" (p. 806), and he recognizes that delusional beliefs call into question how sensitive to evidence beliefs need to be. Nevertheless, he says that "most of our beliefs are sufficiently responsive to evidence to justify the attribution of (literally) innumerable beliefs to us" (p. 806). True as that may be, I think it's the marginal cases that we need to focus on, since the question is whether implicit biases are located at the margins of belief. And once we look at the marginal cases, the cases where we have, for example, beliefs that we might not be aware of, that are insulated from our other beliefs, that don't update in the right sort of way, or that we wouldn't endorse on reflection, it's hard to know how to distinguish them from implicit biases. Again, this doesn't mean that implicit biases have a lot in common with paradigm examples of beliefs. It's that paradigm examples of beliefs don't have a lot in common with our other beliefs, which makes it hard to know why implicit biases shouldn't be included among them.

There is, I think, a general lesson to draw from all of this. What the existence of these marginal beliefs suggests to me is that distinguishing implicit biases from beliefs is going to

involve an incredibly difficult balancing act. It'll require us to find some property that all beliefs have and that implicit biases lack. But I doubt that such a property exists. First, it's enough of a challenge to find necessary and sufficient conditions for most things, let alone for phenomena as complex and seemingly heterogenous as beliefs. Second, since these marginal cases of belief have so many salient features in common with implicit biases, it's doubtful that we'll be able to find any one feature that distinguishes the two. So, whether we focus on evidence-sensitivity or some other property that all and only beliefs are thought to have, I suspect we'll find either that implicit biases can have it too or that some beliefs lack it. Of course, I'm only making a bet here. A more thorough defense of this claim would have to proceed on a property-by-property basis, and there could be, on the final analysis, a property that separates even marginal cases of belief from implicit biases. But when we take into consideration the fact that beliefs can have all of the properties that seem to make implicit biases theoretically interesting, namely, that they too can be unconscious, recalcitrant, and poorly integrated with the rest of our attitudes, I think it starts to seem like a smart bet to make.

So far, I have focused only on implicit biases. That's no mistake, since I think more than the other belief attributions that I wanted to make in the cases above, the claim that Andy believes that Black men are likely to commit crimes (and thus takes this to be a reason to believe that the man holding the purse probably stole it) is highly controversial, as Gendler's and Levy's worries make clear. However, it's worth noting that the view that I want to defend doesn't rest on the claim that implicit biases involve beliefs. Really, what I want to defend is the general claim that the kind of representations that model-based learning systems use to navigate environments count as beliefs (and so, when those states guide our attitudes and actions, we have sufficient grounds to say that we did whatever it was that we did because we recognized reasons

to do it). I also happen to think that model-based learning systems play an integral role in implicit biases, and because of this additional commitment, I'm inclined to think that implicit biases involve beliefs as well. Still, as long as you find it plausible to think that, in the Obama case or in the Kindness case, the agents have the beliefs that I want to ascribe to them, I think I've done enough to show that the recognition view has the resources to say that, even though we often do things automatically, we still do them for reasons.

6. Conclusion

Let's take stock of what I've argued up until now. In the last chapter, I argued that we have good reasons to look for an alternative to the reflection view. In this chapter, I argued that at least some versions of the recognition view allow us to accept that our attitudes and actions often issue from automatic processes and still maintain that we often do things for reasons anyway. In particular, I've tried to show that as long as we accept (1) that our everyday exercises of agency draw on a broad cluster of practical and social skills, (2) that skills are often supported by model-based decision-making systems, and (3) that the kinds of information structures that systems use to help us make decisions support ascriptions of normative beliefs, then at least in a lot of cases when we act on these systems, we can be said to do things because we believe that we have reasons to do them.

However, as I've already made clear, these model-based decision-making systems aren't perfect. Indeed, there's good reason to think that they're heavily implicated in implicit biases. So, I take it that being a competent human agent requires more than just rational guidance. It also requires us to reflect on the reasons for which we do things and to try to do things for better reasons if those reasons don't seem very good to us. Thus, while I don't think this kind of

reflective control is necessary for rational guidance, I still think it plays an important regulative role in exercises of human agency.

Still, if a lot of what we do is the product of implicit model-based processes, as I've tried to argue, then to what extent can we exercise this sort of control over our attitudes and actions? After all, wouldn't it require to know what the reasons for which we do things are? And isn't that precisely the sort of self-knowledge that we often seem to lack? What's more, even if we do know what the causes of our attitudes and actions are, research suggests that we still aren't likely to use reflection to question our attitudes and actions. Instead, it seems like we're more likely to use it to rationalize them. So, again, why should we think that reflection can play this regulative role? In the next two chapters, I turn from questions about rational guidance and try to address these worries.

Chapter 4

Reflection Without Introspection

In this chapter, I want to defend the idea that we can use reflection to monitor and control the processes behind our attitudes and actions against the worry that we often lack access to those processes. More specifically, I want to defend two main claims: (1) that we often know, albeit through *indirect* means, what the causes of our attitudes and actions are and (2) that reflection doesn't require us to have direct access to its objects. So, given both of these claims, I think that we often are in a good position to use reflection to monitor and control the reasons for which we do things.

1. Doris's Skeptical Challenge

There have been a number of philosophers who have recently claimed that our apparent lack of self-knowledge is a problem for reflectivists views of human agency, but the most forceful version of this challenge comes from John Doris (2015). [1] Briefly put, Doris's argument goes as follows:

> P1. We can't exercise human agency unless we would recognize the causes of our actions as reasons for those actions were we aware of them.
>
> P2. For all we know, we wouldn't recognize the causes of our actions as reasons for those actions were we aware of them.
>
> C. So, for all we know, we don't exercise human agency.

Ultimately, Doris wants to avoid this skeptical conclusion, but because he thinks that (P2) is defensible, he concludes that the problem lies with (P1). What's more, since he assumes that all

[1] For example, see also Nahmias (2007) and Kornblith (2012).

reflectivists are committed to (P1), he thinks that this argument gives us a reason to reject reflectivism.

But why should we accept (P2)? This, I take it, is the key empirical premise, and it's based on two further observations. First, Doris assumes that many of the stimuli found in the studies that I discussed in the first chapter are arbitrary. Second, he assumes that since the participants in these studies often aren't aware of these effects when they occur, we have good reason to doubt whether we're aware of them outside of the lab. Thus, according to Doris, we're typically not in a good position to know whether our actions have arbitrary causes, which, given (P1), means that we're usually not in a good position to know whether we exercise human agency, at least as reflectivists conceive of it.[2]

In the last chapter, I argued that the stimuli used in research in social psychology might not be as arbitrary as they seem, and I take it that we could reject (P2) on these grounds alone. But for the sake of argument, let's grant that the causes of our attitudes and actions might often be arbitrary. What I want to focus on is how often are we aware of such causes outside of the lab.

To begin, I think it helps to consider Nisbett and Wilson's (1977) view. In that paper, their main claim is that we typically lack introspective access to the processes behind our attitudes and actions. But they don't claim that we lack all kinds of access to those processes. In fact, they say that we will often be right about the causes of our actions; it's just that, on their view, we usually acquire this knowledge indirectly, by developing causal theories about

[2] Doris isn't very clear what the scope of his skeptical argument is. Sometimes the worry seems to be that we're never in a position to know whether the causes of our actions are arbitrary, but at other times he seems to want to defend only the weaker claim that we very often aren't in a position to know this. Since I want to argue that we have good reasons to reject even this weaker claim, that's the version of his argument on which I'll focus.

ourselves in the same way that other people do, through observation of and inference from our own behavior. Hence, they write that:

> [W]e will often be right about the causes of our judgments and behavior. If a stranger walks up to a person, strikes him, and walks away, and the person is later asked if he likes the stranger, he will reply that he does not and will accurately report the reason. The interaction he has had with the stranger will be highly salient and a highly plausible reason for disliking someone. And, in general, the conditions that promote accuracy in verbal report will be the opposite of those described previously. These conditions may be summarized briefly by saying that reports will be accurate when influential stimuli are (a) available and (b) plausible causes of the response, and when (c) few or no plausible but noninfluential factors are available. (p. 356)

In its broader context, the point of this passage is to highlight an important epistemic difference between participants in studies in psychology and people outside the lab. As Nisbett and Wilson note earlier (p. 350), studies in psychology often trick participants by using stimuli that are either difficult to detect or not very plausible causes of the attitudes and actions that they influence, and so the participants often aren't in a good position to develop accurate causal theories. But Nisbett and Wilson deny that is the case outside of the lab; there, as they tell us above, the causes of our attitudes and a ctions are often available to us and plausible. Thus, according to Nisbett and Wilson, unlike participants in these studies, we will often be right about the causes of our attitudes and actions.

If Nisbett and Wilson are right, then we can start to see a problem with the second premise of Doris's argument. He claims that we should doubt the accuracy of our own self-reports about reasons precisely because people are often wrong about the causes of their attitudes and actions in experimental settings. But this claim ignores the fact that the routes to self-knowledge available to participants in these studies are often much more constrained than they are in our everyday lives. So, just because they're often ignorant of the causes of their action, it doesn't follow that we are too.

However, Doris might respond to this objection in one of two ways. First, he might insist that the conditions of self-knowledge that Nisbett and Wilson identify don't obtain very often. So, despite what they seem to think, we often aren't in a position to know what the causes of our actions are.[3] Second, Doris could also claim that even if Nisbett and Wilson are right that we often have indirect knowledge of the causes of our actions, this still won't help out reflectivists. That's because, according to this response, reflection requires us to have introspective access to its objects. Thus, given that Nisbett and Wilson's main point is that we typically lack such access to the processes behind our attitudes and actions, that shows that we typically can't use reflection to improve our agency.[4]

What I want to do in the rest of this chapter is to show why neither of these responses works. I'll begin, in the next section, with the claim that we often don't know, whether directly or indirectly, what the causes of our attitudes and actions are. After that, I'll turn to the question of what reflection requires.

2. Self-knowledge without introspection

In Doris's defense, I'm willing to admit that the conditions of self-knowledge that Nisbett and Wilson identify in the quotation above (namely, that the causes of our attitudes and actions have to be (a) salient, (b) plausible, and (c) at least as plausible as other available factors) might not obtain all that often. It's one thing to say that we usually know why we're angry when someone punches us, but in most cases, the causes of our attuites and actions aren't literally hitting us in the face. So, if we're going to defend the claim that we often know what the causes of our attitudes and actions are, we would need more routes to self-knowledge than this.

[3] See Doris (2015; pp. 146-47).
[4] Doris suggested this response in conversation.

Still, there's no reason to think that Nisbett and Wilson are identifying the only conditions of self-knowledge. Surely, we can learn about the causes of our attitudes and actions in lots of different ways. In some cases, we might acquire this knowledge through observations of and inferences of our own behavior, but in other cases we might acquire it by talking to friends or therapists or by reading research in social psychology. So, the question that I think we should ask here isn't the narrow question of how often do the conditions that Nisbett and Wilson identify obtain. Rather, it's how often do we know, in one way or another, what the causes of our attitudes and actions are.

And once we adopt a more inclusive view about the routes to self-knowledge, I think it becomes a lot easier to show that, despite what Doris thinks, we often do know what the causes of our attitudes and actions are. The reason why is simple: the more routes to self-knowledge available to us, the more opportunities we'll have to acquire this knowledge. Granted, this doesn't mean that we actually do often know what the causes of our attitudes and actions are – you can have lots of ways to get somewhere and still never get there. But I do think it makes the task of showing that our self-ignorance isn't nearly as pervasive as Doris thinks much more manageable.

So, given that we can learn about the causes of our attitudes and actions in lots of different ways, how often do we have this self-knowledge? I admit that questions about frequency can be hard to answer. How often is often enough to convince a skeptic like Doris that we can have some faith in our self-reports about reasons? Do we have to be right 50% of the time? Or is it 51%? And how do we even go about showing that we're right this often anyway, whatever the relevant frequency is?

Admittedly, I'm not sure how to answer these questions. But I think we can still make progress in this debate by focusing on some of the effects that Doris himself cites and asking whether we're as ignorant of these effects as he assumes. And if it turns out that we're not, then maybe we have good reason to be a more sanguine about the prospects of our self-knowledge in general.

Unfortunately, though, social psychologists haven't spent much time exploring the extent to which people are aware of these effects outside of the lab, and so in most cases we can only speculate about what the answer might be. But there's one exception here, and it's research on implicit biases. Indeed, because of the important political, moral, and theoretical issues that these biases raises, researchers have conducted lots of studies trying to figure out the best ways to regulate their effects on us, and one of the first questions that they have had to address whether (and to what extent) people are aware of their biases and the conditions under which they're apt to be triggered (after all, we can't be expect to regulate the effects of these biases if we don't know that we have them). And since Doris himself cites implicit biases as an example of the kind of arbitrary cause that he thinks we're often ignorant of, this body of research is directly relevant to his argument.

So, what does this research show? The answer, in short, is that not only do most people seem to be aware that they have implicit biases, but they also seem to have a fairly accurate sense of the conditions under which their biases are likely to be triggered. Most of the data here come from Margo Monteith and her colleagues, who've developed and tested a way to measure people's awareness of their biases.[5] They call them "Would/Should Discrepancy Questionnaires," and they work by asking people to report how they *would* and how they *should* respond to various bias-related stimuli. For example, one of the questions on the race-based

[5] For the most recent review of this research, see Burns, Monteith, and Parker (2017).

version of this questionnaire asks the participants to rate, on a seven-point scale, the extent to which they agree with the statement, "I would feel uncomfortable shaking a Black person's hand," where a 1 would indicate that the person strongly disagrees with this statement and a 7 would indicate that they strongly agree with it. Then, on the next question, they're asked to do the same thing, except this time the statement reads, "I should feel uncomfortable shaking a Black person's hand." Once the participants have filled out the questionnaire, which usually contain around 16 different bias-related scenarios, Monteith and her colleagues then compute their "discrepancy score" by subtracting the *should* rating from the *would* rating on each question and adding up the resulting differences. So, for example, if my *would* rating on each question was 5, and my *should* rating was consistently 1, then my overall discrepancy score would be 64. Notice, however, that since two people with different normative outlooks can still end up with the same discrepancy score (e.g., someone who routinely lives up to their egalitarian ideals could have the same discrepancy score as a committed racist), these questionnaires aren't a measure of people's prejudices. Instead, they're a measure people's awareness of bias-driven mismatches between their behavior and explicit attitudes.

Since they started doing this research in the mid-90s, Monteith and her colleagues have given these questionnaires to hundreds of participants, and what they have consistently found is that most of them (approximately 80%) have positive discrepancy scores, meaning that the vast majority of the participants who fill out these forms report that they're prone to bias-driven discrepancies in their behavior across a variety of situations (Monteith *et al.,* 2017). Perhaps even more surprisingly, this finding holds regardless of the participant's age, gender, income, and education level (Monteith & Voils, 1998). So, if we take these data seriously, it's not just

college students who are sensitive to their biases. It's lots of people from many different walks of life.

Of course, as Monteith and her colleagues (2017) themselves recognize, there are good reasons to think that we shouldn't take these data too seriously. In particular, it's reasonable to doubt whether these questionnaires really are an accurate measure of people's awareness of their biases. For one thing, they rely entirely on self-reports, which might be especially dubious in the contexts of implicit biases – not only because these biases operate unconsciously, but also because there might be social pressure on participants to over- or under-report the extent to which they're susceptible to their effects. Moreover, by forcing people to think about situations in which biases are likely to influence our behavior, these questionnaires might also make those situations more salient than they would normally be. So, just because someone says that they would respond in a biased way in those situations, that doesn't mean that they are disposed to notice their biases in "real life."

In response worries, Monteith and her colleagues have conducted studies to test the accuracy of these questionnaires, and the results are encouraging. For example, in one of those studies, Monteith and Voils (1998) found that discrepancy scores have some predictive value: the participants who said that they were less prone to bias-driven discrepancies in their behavior really were less prone to them, particularly when they were put in conditions of high cognitive load, which had already been shown to make people more suspectable to the effects of implicit biases (Wigboldus *et al.*, 2004). Similarly, in another study, Monteith, Mark, and Ashburn-Nardo (2010) conducted one-on-one interviews with over 150 White participants in which they asked them, among other things, whether they could recall "racial experiences in which they thought, felt, or acted in a way toward African Americans that they later wished they had not" (p. 235),

and they found that discrepancy scores were positively correlated with not only the number of such experiences that the participants could recall but also the extent to which they seemed to regret their behavior during those experiences. So, in other words, the higher someone's discrepancy score was, the more they appeared to be sensitive to and aware of situations in which they acted in biased ways.

Just to be clear, I don't want to suggest that these questionnaires are perfectly accurate. In all likelihood, they're not. Nor do I want to say that people are always aware of their implicit biases. But I do think that this research serves as a nice counterbalance to all of the studies that are designed to establish the existence and scope of our implicit biases. Indeed, if all we did was look at this research, we might think that most people are ignorant of their implicit biases and their effects on them. But the studies by Monteith and her colleagues suggest that this isn't obviously true. When they're asked, most people say that they are prone to bias-driven discrepancies in their behavior, and it seems like we have some reason to take them at their word.

But these are *implicit* biases that we're talking about. They're supposed to operate unconsciously. So, how do most people know about them? The answer, according to Monteith and her colleagues, has to do with discrepancy detection: in one way or another, people are able to notice discrepant thoughts and actions, and when they do, they're likely to start looking for the cause of the discrepancy so that they can avoid it in the future. To illustrate how this process works, Monteith *et al.* (2017) give us the following example (which you might recall from the last chapter):

> Imagine that Andy, a White man, is walking through a crowded mall. Like many people, Andy thinks of himself as a tolerant, fair, and egalitarian person. As Andy is walking, he happens to see a Black man holding a purse. The thought, "I bet he stole that purse" automatically enters Andy's mind. Just then, he sees a woman

> emerge from a store, take the purse, and put it over her shoulder. Suddenly, Andy realizes that he stereotyped the Black man as a criminal. Andy might wonder why he had this negative thought. Why did he assume this man was a thief when he was just standing there? Would Andy have responded the same way if the man was White? He feels guilty, and the natural activity of the BIS [i.e., the Behavioral Inhibition System] causes him to momentarily pause and note stimuli in this situation (e.g., his location, what he was doing, the race of the other person, etc.). Upon encountering these cues in the future when a similarly biased response may occur, Andy's BIS will likely become activated and he will be able to inhibit biased thoughts while maintaining an egalitarian mindset. (p. 11)

There's a lot going on in this passage, but what I want to focus on are the early parts of the process, especially the part where Andy comes to realize that he stereotyped the Black man. Although Monteith and her colleagues don't have much to say about how discrepancy detection happens, I don't see why we should think that it always involves the same kind of process. In fact, if we adopt an inclusive view about the routes to self-knowledge, then any route to this information would seem to do the trick. In the example above, I take it that Andy learns about his racist thoughts through a combination of introspection, observation, and inference: presumably he has direct access to the content of his thoughts, but it's only once he sees the woman walk out of the store and take her purse that he realizes that he was mistaken to assume that the Black man probably stole it, which in turn causes him to wonder why he even made this assumption in the first place and whether or not it was motivated by bias (e.g., "Would I have responded the same way if he was White?"). All of this strikes me as quite plausible, but notice that there are other ways the story could have gone as well. Here are just a few of them:

1. Andy is walking through a crowded mall with his good friend Annie. Like many people, Andy thinks of himself as a tolerant, fair, and egalitarian person. As he and Annie are walking, they happen to see a black man holding a purse, and Andy automatically turns to Annie and says, "I bet he stole that purse." Without hesitation, Annie shoots him a look of disapproval. "Why?" she asks. "Because he's Black? If he were White, would

you have said the same thing?" At first, Andy is defensive; he denies that his comment was motivated by bias and eventually tells Annie to just forget that he even said it. But later that day, as Andy finds himself thinking about what happened some more, he comes to realize that Annie was probably right. Because of his own biases, he stereotyped the Black man as a criminal.

2. Andy is walking through a crowded mall alone. Like many people, Andy thinks of himself as a tolerant, fair, and egalitarian person. As he is walking, he happens to see a security guard confront a group of Black teenagers, and out of curiosity he decided to eavesdrop on their conversation. Things start to get heated pretty quickly, and at one point he hears one of the teens say to the security guard, "Would you accuse us of stealing if we were White?" The guard refuses to answer this question, but as the exchange continues, it becomes clear to Andy that he has no real evidence that they committed a crime (aside, that is, from the fact that they fit the profile of someone whom the guard would expect to shoplift). Andy is initially outraged by the guard's behavior, but then he starts to wonder what he would do if he were in the guard's situation. They are, after all, quite similar — same age, race, gender, and they probably grew up in similar neighborhoods and have similar families. Would he have profiled the teens too? Just then, Andy remembers something that happened last week, when he was at the very same mall: he saw a Black man holding a purse and mistakenly assumed that he stole it. "Would I have made this assumption if he were White?" he thinks to himself. "Maybe I'm not very different from the guard after all — maybe I would've profiled the teenagers too."

3. Andy is walking through a crowded mall alone. Like many people, Andy thinks of himself as a tolerant, fair, and egalitarian person. As he is walking, he happens to see a Black man holding a purse and automatically thinks to himself, "I bet he stole that purse." Just then, he remembers an article that he recently read online about implicit biases. In it, the author explained that most people have prejudices that they might not be aware of and that can cause them to do, think, or feel things that conflict with their values. Andy starts to wonder, "Is this one of those situations that the author was talking about? Would I have assumed that this guy stole the purse if he were White?" After thinking about it just a bit, Andy realizes that his thought was probably motivated by one of these biases.

In all of these cases, Andy comes to learn about his biases in different ways. In the first case, it's primarily through testimony; Annie tells him that his behavior was motivated by bias, and he eventually comes to agree with her. In the second case, it's through analogical reasoning; he believes that the security guard's behavior is motivated by biases, and this causes him to reflect on what he would do in similar circumstances. Finally, in the last case, it's through induction; because he recently read that article, Andy knows some general facts about implicit biases, and he applies this knowledge to his own situation. From what I can tell, these are all perfectly familiar and indeed legitimate ways in which we learn about ourselves, and yet none of them requires us to have introspective access to the processes behind our attitudes and action. So, if we want to know how so many people could know about their implicit biases, here's a plausible answer: they learn about them by using the very same methods that everyone else uses.

3. Generalizing the Findings

If Monteith and her colleagues are right and most people do have at least some awareness of their implicit biases, then I think there are a couple broad lessons that we might draw about Doris's arguments. First, just because a process is automatic, we can't assume that we lack access to it. We might lack *introspective* access to it, but introspection is far from the only way that we can learn about ourselves.

Second, although much more research is needed, I'm inclined to think that Monteith and her colleagues' findings might generalize to our awareness of other kinds of discrepancies between our actions and personal standards as well. Of course, there are good reasons to think that people might be especially concerned about their prejudices (either because they really are committed to egalitarian ideals or because they simply want to avoid the social sanctions that are associated with racist or sexist behavior), but we have plenty of motivation to avoid other kinds of discrepancies. As a result, I suspect that once we notice them, we'll often want to know what their causes are too.

To see what I have in mind, consider a case of weakness of will. In the standard case of weakness of will, the agent is aware of what she's doing and why she's doing it. So, for example, when you give into temptation and opt for that second slice of cake even though you just told yourself that, all things considered, you really shouldn't, you know exactly what you're doing (eating a slice of cake) and why (because you want to). However, to bring the example a little closer to home, let's imagine that this isn't the case — the agent doesn't know, at least not right away, what she's doing or why. So, for example, imagine that I'm trying to lose weight, but after a few weeks of dieting, I start to fall into subtle patterns of overeating. Let's also suppose that, because of how subtle these patterns are, I don't realize that I'm doing this. In fact, if you were to

ask me, I'd tell you that I'm adhering to my diet pretty well. But, of course, I'm wrong, and if I were to find out that I've been overeating, it certainly wouldn't make me very happy.

Now, it seems to me that, even though my eating habits aren't directly available to me on introspection, there are still lots of ways that I can learn about them. Maybe I step on a scale a few weeks later and notice that I haven't been losing weight at quite the rate that I would like, or maybe someone else notices my eating habit and points them out to me. In any event, it's not a deep mystery how this could happen, and once it does, it's highly plausible to think that I'll start to wonder why I've been overeating. Is it because I've been under an unusual amount of stress recently? Or maybe it has more to do with my environment. Or maybe it's some other factor. Whatever the case may be, since I'm already motivated to lose weight, this is information that I want to have, and although it might not be easy to figure out what the cause of my overeating is (sometimes it takes hours of therapy to figure out why we do the things we do), I see no reason to think that I wouldn't be able to eventually come up with a perfectly good answer.

All of this is, admittedly, quite speculative, and like I said, much more research needs to be done before we can know the extent to which people are aware of discrepancies in their thoughts and actions. I also admit that such discrepancies can be difficult to detect, and that's in no small part because we don't always get the right kind of feedback on our attitudes and actions. Indeed, a major reason why the effects of our implicit biases often go unnoticed is that we don't have good opportunities to learn that we're acting in biased ways (things are much different, no doubt, for the people who are on the receiving end of this behavior). Consider, for example, the famous study by Bertrand and Mullainathan (2003) in which they found that job applications were much more likely to get callbacks if they had a White-sounding name than if they had a Black-sounding name. Now, when we have the data in front of us, these hiring patterns are easy

to identify, but it's not surprising that the employers wouldn't be aware of any biases in their practices. After all, they're not privy to these large-scale statistical trends. That's not to say that this information is totally unavailable to them. Maybe they could figure out that their practices are biased by looking around the office and seeing a sea of White faces. Still, in many cases, discrepant hiring practices aren't very easy to detect, and that's because employers aren't receiving the feedback needed to know that they're acting in biased ways. But what I think this shows isn't that people tend to be completely ignorant of their implicit biases, but that they're likely to have certain blind-spots or gaps in their self-knowledge, especially when they lack opportunities to notice discrepancies in their thoughts and actions.

All told, then, I take it that a basic assumption behind Monteith and her colleagues's model of how we learn about our implicit biases is that it's in virtue of the fact that people want to avoid to discrepancies between their attitudes and actions and their personal standards that they're inclined to search for the causes of these discrepancies once they notice them. If so, then one prediction that their model makes is that people will often be aware of the causes of their attitudes and actions when they're in a good position to notice such discrepancies. Obviously, though, this doesn't mean that we're always going to be in this position or that, when we notice one of these discrepancies, our attempts to figure out its cause are always going to succeed. But I do think that, if Monteith and her colleagues are right, our self-ignorance isn't nearly as pervasive as Doris thinks.

4. Reflection without introspection

But even if we do have a lot of indirect self-knowledge, you might still wonder what any of this has to do with reflection. After all, there seems to be a close connection between reflection and introspection such that we can't reflect on something unless we have

introspective access to it. Doris himself is quite clear about this connection when he writes that "[r]eflection can be understood as '[t]hinking about one's own mental processes from a first-person point of view (Kornblith 2012; 28),' which looks a lot like what often gets called introspection" (p. 19). So, given that I haven't tried to defend introspection, you might think that I haven't defended reflection either.

But it's not obvious that reflection always involves introspection. Indeed, it's natural to think that we can reflect on things like our character traits and other behavioral tendencies. For example, maybe I'm not as honest as I would like to be, and so I start to think about why and what I can do to improve my character. In that case, it sure seems like we can describe what I'm doing here as a kind of reflection, and yet it also seems clear to me that its object (i.e., my lack of honesty) isn't something that I can learn about through introspection alone. To be sure, on any given occasion when I lie, I might know what I'm doing through introspection. But knowing that I am lying here and now isn't the same thing as knowing that I have a tendency to lie. The latter seems to require additional operations (e.g., an inference from observations of my own behavior), and so it's not the sort of thing that we can directly know about ourselves. And if that's right, then it would seem to follow that reflection doesn't require us to have introspective access to its objects. In other words, it seems like there are things that we can reflect on that we learned about in other ways.

Now, there are two things that one could say in response to this argument. First, it could be denied that we can reflect on things like our character traits and other behavioral tendencies. But this claim seems highly revisionary, and at any rate I assume that most reflectivists would deny it. So, unless the debate between Doris and reflectivists turns out to be a verbal dispute on what "reflection" means, I take it that we should stick with standard usage and continue to think

of reflection as the sort of process that can be directed at our character traits and other behavioral tendencies.

But there's another way to respond to this argument, which is to deny that we can't learn about our character traits and other behavioral tendencies through introspection. At times, Doris seems to be open to this idea. For example, right after the quotation above in which he tells us how he understands reflection, he writes that, "introspection can be interpreted capaciously, to include first person thinking on things not readily described as mental states, such as habits and dispositions" (p. 19). He then goes on to say that, "[w]hether or not this should be counted introspection, it ought be counted as reflection: I might reflect that I habitually study at Kaldi's coffee shop, or that I'm disposed to curt responses when conversing with Edgar" (p. 19). I think it's clear that Doris is trying to be ecumenical here, and so maybe he doesn't have (or at least doesn't need to have) a considered view about what the proper objects of introspection are. But if he does want to count this sort of "first person thinking on things not readily described as mental states" as introspection, he could say that we can reflect on things like our character traits and habits without also giving up the idea that reflection requires us to have introspective access to its objects.

However, there are two things to note about these remarks. First, if Doris really is willing to adopt such a capacious view about introspection, then, for the reasons that I gave in the last two sections, it will be much harder for him to defend his skepticism about agency. So, I think there is a real question here about whether or not Doris is really entitled to such a view, given his skeptical argument.

Second, this sort of "capacious" view of introspection also seems to be quite revisionary. In particular, although there's a great deal of disagreement about what exactly introspection is, I

take it that almost everyone involved in this debate agrees that when we learn about something through introspection, we learn about it in an immediate, non-inferential sort of way (Schwitzbegel, 2010). But it's hard to image that Doris could've learned about his habit of studying at Kaldi's or about his disposition to curt responses when conversing with Edgar in this way. Indeed, it seems like these are precisely the sort of things that he would have to learn about himself by using the same methods that are available to everyone else. So, unless Doris means for knowledge that we acquire on the basis of introspection to simply refer to any knowledge that we have about ourselves (however it's acquired), I don't see how he could learn about these behavioral tendencies on the basis of introspection alone. Instead, it seems to me like other routes to self-knowledge must be involved as well.

To summarize, then, the argument for thinking that reflection doesn't require us to have introspective access to its objects proceeds from three observations, which we might state as follows:

1. Reflection requires us to have introspective access to its objects.
2. We can reflect on things like our character traits and other behavioral tendencies.
3. We lack introspective access to our character traits and other behavioral tendencies.

When taken at face value, each of these claims might seem plausible, but they can't all be true, and so at least one of them has to go. Thus, if we accept (1), then we'll have to reject either (2) or (3). But rejecting (2) requires us to make revisionary claims about the scope of reflection, and rejecting (3) seems to be incompatible with the widely held view that introspection involves a kind of immediate access to its objects. So, I'm inclined to think that the best option here is to reject (1).

But even if you don't buy this argument, I think there are other ways to cast doubt on the idea the reflection always involves introspection. For example, imagine someone — let's call him "Abraham" — who lacks introspective access to his desires but who, as luck would have it, has a direct line of communication with God. One day, God lets Abraham know that he (Abraham) has a desire to eat cake, and this causes him to think about whether he should act on this desire. On the one hand, he thinks to himself, cake is delicious, and so that's a good reason to get some. But it's also full of calories and sugar, and so maybe he shouldn't. In the end, Abraham decides that he shouldn't eat any cake, and he acts accordingly. Now let's ask: does the process by which Abraham arrives at this decision count as reflection? If we assume that reflection requires us to have introspective access to its objects, we would have to say that it doesn't. But that strikes me as implausible. Indeed, aside from the admittedly odd wrinkle about his Abraham learns about his desire, this is a paradigm example of reflective control. So rather than denying that Abraham is engaged in reflection here, I think it's more plausible to say that reflection doesn't always involve introspection.

This is obviously an extremely example, but notice how similar the structure is to another case we've already considered, namely the case in which Andy learns about his biases thanks to Annie's testimony. After all, in both cases, the agent, we can assume, lacks introspective access to a potential cause of his behavior, and he comes to learn about that cause based only on someone else's testimony. And if it's plausible to think that Abraham can reflect on his desire to eat cake, I see no reason to deny that Andy can reflect on his biases as well.

5. Conclusion

In this chapter, I defended two main claims. First, I argued that, despite what Doris thinks, we often do seem to know, albeit indirectly, what the causes of our attitudes and actions

are. Second, I argued that reflection doesn't require us to have introspective access to its objects. And if both claims are right, it follows we're often in a good position to reflect on the causes of our actions.

But, of course, none of this to say that we actually do engage in this kind of reflection. Instead, all I've taken myself to show in this chapter is that, if we were sufficiently motivated to use reflection to improve our agency, we often could. But how often are we actually sufficiently motivated to use reflection in this way? That's the question that I want to address in the next chapter.

Chapter 5

Reflection with Other People

Suppose that we have all the self-knowledge that we need in order to reflect on the processes behind our attitudes and actions. Does that mean that we'll often use reflection to improve our agency? Unfortunately, it doesn't, and there are two separate worries here. First, studies have shown that we tend to lack the motivation needed to reflect critically on our attitudes and actions. Second, even when we are sufficiently motivated to engage in this kind of critical reflection, there's still no guarantee that we'll get things right. In fact, it could be that reflection often makes us worse off. So, why should we think that we can count on reflection to improve our agency?

My aim in this chapter is to show that these problems become much more manageable if we focus on the social aspects of reflective agency. In the last chapter, I highlighted one way that other people can improve our ability to engage in reflection: they're often an important source of knowledge about the causes of our attitudes and actions. In this chapter, I want to highlight two other ways that they can do this: (1) they're often an important source of motivation to engage in critical reflection, and (2) they can often help us to correct biases that would interfere with our reasoning. So, as I see it, the key to reconciling the idea that we can use reflection to improve our agency with research on motivated reasoning is to stop thinking about reflective agency as an individual achievement and to start thinking of it as something that crucially depends on other people.

1. Background Assumptions about Reflection

But before I defend these claims, I want to lay out a few background assumptions about reflection. First, I don't assume that people should always reflect on their attitudes and actions.

In part, this is for psychological reasons: because we have only so much time and energy, it makes sense that we would engage in reflection only when we take ourselves to have good reasons to do so. However, it's also for moral reasons: in some situations, it seems like we don't want to encourage people to engage in reflection. For example, if someone you care about is drowning, you shouldn't have to reflect on what you ought to do. Instead, it seems like you should just jump in and save them. So, we need to be careful about how often we engage in reflection. Otherwise, we might deplete our cognitive resources or find ourselves having "one thought too many" (Williams, 1976).

So, when should we expect people to engage in reflection? Following Valerie Tiberius (2012), I have two different kinds of cases in mind. The first kind of case involves conflicts between our explicit attitudes. For example, imagine someone who enjoys watching football on Sundays but who also disagrees with the NFL's response to domestic assault allegations, on-field protests, and research on concussions. So, they're conflicted: on the one hand, they want to maintain their weekend routine but, on the other hand, they have serious moral concerns about the NFL. In that case (and assuming that they're aware of this conflict), I think we can expect them to use reflection to help resolve it. Is there a way to reconcile their desire to watch football on Sundays with their concerns about the NFL's policies? Or do they need to either adjust their weekend routine or make certain moral allowances? These are questions that I think reflection can help them to answer.

The other kind of case that I have in mind involves conflicts between our explicit attitudes and our automatic responses. For example, consider Andy again. Once he realizes that there's a discrepancy between his egalitarian values and his belief that the black man stole the purse, he has an opportunity to reflect on whether he should have this belief, what its causes

might be, and what he can do to avoid such thoughts in the future if he decides that they're unreasonable. Likewise, I might explicitly believe that morality is impartial but often find myself acting on certain tribalistic tendencies. In that case, as long as I take these tendencies to be incompatible with my moral commitments, I think we can again expect me to engage in reflection – where do these tendencies come from, are they reasonable, and what can I do to change my behavior if I decided that they're not reasonable?

In general, what both of these kinds of cases have in common is that, by virtue of noticing that there's a conflict either between their explicit attitudes or between their explicit attitudes and their automatic responses, the agents come to realize that they have reason to change their minds or behavior. What's more, I think it's plausible to assume that the need to make such changes will often lead to reflection. So, on my view, we can generally expect people to engage in reflection when they explicitly take themselves to have some reason to adjust their attitudes or actions. There might be other situations in which we take ourselves to have such reasons beyond the kinds of cases discussed above, but I suspect that they're the most common kinds of cases in which this occurs.

But I also don't assume that reflection will always lead us to make the right adjustments. Obviously, we're fallible, and our ability to engage in reflection is no exception. So, there will be times when, on reflection, we decide to change our attitudes or actions, but the changes we make are misguided. For example, it could very well be that, because her desire to continue watching football on Sundays is so strong, the football fan is able to convince herself that the NFL's behavior isn't that bad after all. So, she decides to resolve the conflict by adjusting her moral beliefs rather than her Sunday routine, but we might think that those weren't the right adjustments to make.

However, I assume that anyone who thinks that we can use reflection to improve our agency is committed to the claim that, at least under the right conditions and perhaps over a long enough period of time, it tends to lead to true beliefs about what our attitudes and actions ought to be.[1] In this sense, we can say that they're committed to thinking that such reflection is conducive to truth. So, if it were to turn out that reflection isn't truth-conducive in this sense, then I take it that we would have good reason to reject the claim that we can rely on reflection to improve our agency.

2. The Motivation Problem

But are people really inclined to use reflection to question their own attitudes and actions? Haidt (2001; pp. 820-21) claims that they not, and his argument, as I understand it, rests on two claims. First, he tells us that reasoning is usually driven by one of two different kinds of motives, which calls "relatedness" and "coherence" motives. Second, he assumes that when our reasoning is driven by either of these motives, we won't be inclined to use it to question our attitudes and actions. Thus, according to Haidt, we typically aren't inclined to use reflection to question our attitudes and actions.

However, I think we can question both of these claims. Let's start with the claim that our reasoning is typically driven by relatedness or coherence motives. By "relatedness" motives, what Haidt means are "concerns about impression management and smooth interactions with other people" (p. 820). For example, studies have shown that we are likely to have a more favorable attitude toward people with whom we expect to interact, and one explanation for this effect is that our desire to get along with them causes us to selectively recall reasons for thinking

[1] Of course, there are complicated questions here about whether beliefs about what we ought to do or think are truth-apt in the first place and, if they are, exactly what are the conditions under which they're true. However, a full discussion of these issues falls outside of the scope of my project. So, I'll just assume that some of these beliefs are true.

that they're likeable (Kunda, 1999). Likewise, studies have shown that our attitudes on various issues are likely to shift toward the views of people with whom we expect to discuss those issues, and again it's plausible to think that these shifts are driven by a desire to get along with others (Chen, Shechter, & Chaiken, 1996).

Coherence motives, by contrast, include "a variety of defensive mechanisms triggered by cognitive dissonance and threats to the validity of one's cultural worldview" (p. 820). For example, as Haidt notes, research on cognitive dissonance has long shown that people will update their beliefs in seemingly irrational ways so that they can make them consistent with their behavior (see, e.g., Festinger, 1957), and studies on the so-called "my-side" bias suggests that we have a selective tendency to seek out, recall, or readily accept evidence that counts in favor of our beliefs and to ignore, forget, or try to discredit evidence that goes against them (see, e.g., Lord, Ross, & Lepper, 1979).

But these aren't the only motives that can influence our reasoning. We're also sometimes motivated to engage in reasoning simply because we want to arrive at true beliefs, and studies have shown that such "accuracy goals" can improve our reasoning. For example, Tetlock and Kim (1987) found that when people are motivated by accuracy goals, their reasoning tends to be more systematic and elaborate, which in turn increased the accuracy of their judgments on a variety of tasks. Likewise, other studies have shown that accuracy goals can make us less susceptible to a number of cognitive biases, including fundamental attribution errors (Tetlock, 1985), anchoring effects, primacy effects on impression formation, and ethnic stereotyping (Kruglanski & Freund, 1983). What's more, subsequent studies suggest that the reason why many of these effects go away is that accuracy goals improve our reasoning. Indeed, when

participants weren't given as much time on a task and thus couldn't reason as carefully, many of these biases returned (Freund, Kruglanski, and Shpitzajzen, 1985).

So, I take it that there's some reason to think that accuracy goals can improve our reasoning. But under what conditions are they likely to be activated, and how often do those conditions obtain? The first question is easier to answer. In particular, it appears that there are two main sources of accuracy goals. The first is accountability: if we think we'll have to justify ourselves to others, we're likely to be motivated by accuracy. In the study by Tetlock and Kim, for example, this is how they motivated participants to be accurate: they told them that they would have to explain their reasoning to researchers afterward. The other source of accuracy goals is related to our concerns: if we think that something important or something that we care about depends on the outcome of our reasoning, accuracy is again likely to be our goal. For example, in some of the other studies above, the participants were told that their judgments could affect other people's lives or that it was important to the experimenters that they get things right, and this also motivated them to be accurate.

However, we can only speculate about what the answer to the second question is. Still, I don't think it's implausible to suggest that these conditions might obtain relatively frequently. After all, it seems like we often do expect to have to justify ourselves to others, and although not every decision we make is consequential, surely we recognize that at least some of them are. So, if these are the conditions under which accuracy goals are apt to be triggered, and if accuracy goals really do improve our reasoning, then you might think we can often use reasoning to improve our agency.

Nevertheless, research on accuracy goals is a bit of a mixed bag. For one thing, they don't always improve our reasoning (Fischhoff, 1977; Kahneman & Tversky, 1973), and that's in large part because the motivation to get things right can take us only so far. We also have to know what the right kind of reasoning strategies are for the task at hand. But this information isn't always available to us, and in those situations, accuracy goals aren't going to improve our reasoning (Kunda, 1999).

Second, accountability doesn't always give rise to accuracy goals. In fact, studies have shown that this is likely to happen only when we don't already know what our audience thinks about the issue at hand. But if we have this knowledge, our reasoning tends to be biased in their favor – we'll try to figure out not what the truth is, but whether there are attitudes or actions available to us that we can justify in terms of reasons that they'll accept (Tetlock, 1983; Telock, Skitka, & Boettger, 1989). So, a lot depends on our audience. If we don't know their views in advance, we're likely to be driven by accuracy. But if we do, we'll probably be driven by relatedness goals instead.

As a result, I don't think the best way to respond to Haidt's argument is to reject the first claim. Research on accuracy goals is just too inconclusive to know whether we're often motivated by these goals and, if so, whether they'll improve our reasoning. So, if that's all that reflectivists can say in response, then maybe we shouldn't think that we often want to question our attitudes and actions.

But I think that reflectivists have good reasons to reject the second claim as well. In particular, it's not obvious to me that research on relatedness and coherence motives shows that people aren't willing to question their attitudes and action. In fact, as long as we have a realistic

view about the conditions under which such reflection is like, I think this research actually supports this claim.

2.1 Coherence Motives

Let's start by focusing on coherence motives. At bottom, I take it that the worry here is that people tend to use reasoning in ways that are ultimately self-serving. Thus, we're not likely to question our attitudes and actions unless we're given some reason to do so, and even then we'll try really hard to defend them anyway. So, as long as reflectivists are committed to thinking that we're the sort of creatures who should always be inclined to question their attitudes and actions and who should always respond to reasons in an "objective" or unbiased way, it looks like they're deeply mistaken.

However, I don't think that reflectivists need to be committed to either of these claims. First, in the last section, I already explained why I don't think that people should always be expected to question their attitudes and actions (namely, that it would be a waste of our limited cognitive resources, and there are good moral reasons to think that we shouldn't always engage in reflection anyway). Indeed, the claim that I want to defend is that we should expect people to engage in reflection when they take themselves to have good reason to do so. So, insofar as research on coherence motives shows that people are willing to engage in reflection when they're given some reason to think that their beliefs are false, it would seem to support my view, not undermine it.

Still, it might be pointed out that the way in which people respond to this evidence is still problematic. After all, it's not like they're approaching the issue of whether they're wrong with an open mind; they are clearly doing everything they can to defend their beliefs. So, it might be

thought that while these studies do show people are willing to engage in some kind of reflection in response to countervailing evidence, it's not the kind of critical reflection that reflectivists associate with human agency. That is, the question that this evidence seems to prompt them to ask isn't, "Is my belief actually false?" Rather, it's something like, "What can I say to discredit this evidence?"

However, note a few things in response to this worry. First, people usually aren't at liberty to believe whatever they want. If the evidence available to them suggests P, and they can't come up with a reason to think otherwise, then they will usually believe that P. As Ziva Kunda (1999) puts it, "we will only draw our desired conclusions when we succeed at justifying them. Our ability to justify desired conclusions is [therefore] constrained by our understanding of reality" (p. 232). So, no matter how much someone might want to defend their beliefs, if they're unable to come up with reasons to discredit the evidence that they're given, chances are they'll end up changing their beliefs (or if they don't end up changing their belief, they'll at least be a little less confident in them).

Second, precisely because researchers know that people are likely to change their beliefs in light of strong evidence to the contrary, the evidence that is used in studies on coherence motives tends to be mixed, meaning that it doesn't unequivocally support one side of the debate over the other (Lord, Ross, & Lepper, 1979). So, it's often the people motivated not to believe the evidence who are clearly responding to it well. They're the ones who are subjecting it to the rational scrutiny it deserves.

Finally, it's not at all obvious that we should expect everyone to respond to evidence in the same way. More specifically, if you have lots of experience that suggests P, and someone

gives you an article that suggests that not P, it's entirely reasonable for you to first question the article, not your belief. Alternatively, if you have lots of experience that suggests that P, and someone gives you an article that supports this belief, it's also entirely reasonable for you to accept its arguments at face value. So, it seems like people's responses to evidence should depend on their prior beliefs.[2]

In short, then, I think any reflectivists should be willing to say that people shouldn't always be expected to reflect critically on their attitudes and actions and that people shouldn't always respond to reasons in the same way. However, I want to make one more point about research on coherence motives, which is that it can be difficult to know what the impact of exposing someone to countervailing evidence is because these studies usually don't examine the long-term effects of such exposure. Instead, they typically focus on people's immediate responses. But our immediate response to countervailing evidence aren't a great measure of its impact on us, since it often takes us a while to come around on an issue. So, if we want to know whether people are willing to change their minds in response to countervailing evidence is, what we need are longitudinal studies.[3]

To see a good example of this, I think it helps to consider yet another series of studies by Margo Monteith and her colleagues (Czopp, Monteith, & Mark, 2006). In these studies, they wanted to know how people would respond to interpersonal confrontations of bias. So, what they first did was put participants in situations where they were exposed to stimuli that were likely to elicit a biased response and had an experimenter who was pretending to be another participant confront them about it if they responded in a biased way. After that, they had the participants

[2] I'm certainly not the first person to make this point. See, e.g., Klayman and Ha (1987).

[3] Obviously, longitudinal studies can be difficult to coordinate, time-consuming, and expensive, so it's not an accident that they're hard to find. Still, I think they could help to answer important questions about the extent to which people are willing to change their minds in response to evidence that goes against their views.

perform another task that was again designed to elicit a biased response. What Monteith and her colleagues ultimately wanted to know was whether or not the confrontation would have an impact on their performance on the second task, and what they found is instructive: no matter how the participants initially responded to the confrontation (that is, whether or not they were open to the criticism or really defensive at first), if they were already motivated not to be biased, then they were much more likely to control the influence that their biases might have on them during the second task. That is, the biggest predictor of whether the confrontation would affect people's behavior over time wasn't their initial response to it. It was their motivation to not be biased in the first place.

On its own, this finding is interesting and important, but I also think it has broader implications. In particular, I take it to show that we shouldn't give too much weight to people's initial responses to criticism when we're trying to figure out whether or not it'll cause them to change their minds. After all, if Monteith and her colleagues stopped the study after the first phase, we might think that many of the participants who were at first defensive in response to the confrontation wouldn't be inclined to change their behavior. But, of course, that's not what they found.[4]

So, I take it that we can expect people to be both more and less open to criticism than research on coherence motives is often thought to show. First, we can expect them to be more open to criticism because people's immediate responses don't always predict the impact that it'll have on them. Second, we can expect them to be less open to criticism because people shouldn't

[4] Some of Haidt's own research is susceptible to this worry. In particular, what I have in mind is his work on moral dumbfounding. Just because many of the participants in this study didn't change their minds immediately after their conversation with him and his colleagues, that doesn't mean it had no effect on them. Perhaps many of them changed their minds later on. Either way, it's hard to know what impact it had on them because all we have are their immediate responses.

always question their beliefs in light off countervailing evidence. So, at the very least there's a real question here about the extent to which this research shows that we're unwilling to question our attitudes and actions.

2.2. Relatedness Motives

Let's now turn to relatedness motives. As I understand it, the basic idea here is that when we're not using reasoning to defend ourselves, we're often using it simply to maintain a good reputation and get along with others. I have no doubt that this is true, but how much should it trouble reflectivists?

If reflectivists were committed to the claim that we can use reflection to improve our agency only if accuracy is our goal, then I think research on relatedness motives would pose a problem for their view. However, I don't think that reflectivists are committed to this claim, and in fact I'm inclined to think that relatedness motives often cause us to reflect critically on our attitudes and actions.

For example, suppose that I tend to leave the cabinet doors open in the kitchen, and it drives my wife nuts. Let's further suppose that, as a matter of fact, I don't think it make a difference whether or not the cabinet doors are closed. So, if I were left to my own devices, it wouldn't even occur to me that this is something that I should do. Nevertheless, because I care about my wife, I reflect on my tendency to leave the doors open and ultimately resolve to make an effort to try to close them. In that case, it seems to me like we could say the following: (1) that I've engaged in the kind of reflection that reflectivists associate with human agency and (2) that it was primarily motivated not by accuracy goals but by relatedness motives. Indeed, my decision to reflect on and regulate this tendency wasn't based on concerns about its reasonableness (as I

stipulated, I don't think that there's anything wrong with tendency). Rather, it's based on a desire to maintain a good relationship with my wife. And if that's right, then reflectivists don't have to claim that accuracy goals are necessary for reflection. They could instead claim that relatedness motives will often do the trick.

But it might be thought that (2) doesn't accurately describe this case. According to this response, what's motivating me to engage in reflection here is a desire to do the right thing, and it just so happens that, in this case, I come to believe that the right thing for me to do is to conform to my wife's preference. In that case, my reflection is motivated by an accuracy goal, not relatedness concerns.

But while this is possible description of what's going on in this case, I don't see why it's necessary. Why couldn't my motivation to reflect on and regulate my tendency to leave the cabinet doors open be driven entirely by a concern to maintain a good relationship with my wife? If it can, then I think we can easily imagine that my motivation here derives from relatedness concerns, no accuracy goals.[5]

I think we find an analogous case by again considering research on the self-regulation of implicit biases. Typically, the participants who want to avoid acting on their implicit biases are divided into two categories. People who are *internally motivated* want to avoid acting on their biases because of their moral commitments. People who are *externally motivated*, by contrast, want to avoid acting on their biases because they don't want to be subject to the social sanctions

[5] Alternatively, if it turns out that the best way to describe this case is in terms of an accuracy goal, there's a case to be made for thinking that many of the studies on relatedness motives might actually implicate accuracy goals as well. For example, in the study that shows that people will often shift their attitudes to match the views of their interlocutors, we could say that they're motivated to do the right thing, which they just so happen to think is to try to be agreeable to strangers. In that case, I would argue that we're motivated by accuracy goals much more often than Haidt and others think. However, since I assume that we don't have to explain these cases by appealing to accuracy goal, my main claim instead is that relatedness motives can often cause us to engage in the sort of critical reflection that reflectivists care about.

associated with biased thoughts or actions. Since external motivation is closely tied to concerns about impression management and avoiding reputational costs, it's plausible to construe it as a relatedness motive, and yet when people who are externally motivated notice that they're acting in biased way, they're often no less likely to reflect on and regulate the causes of their behavior (Monteith, et al., 2017).[6]

So, part of what I want to say in response to Haidt is that we shouldn't assume that relatedness concerns can't motivate us to reflect critically on our own attitudes and actions. But I also suspect that they might often have this effect. This is because, just in virtue of being social creatures, we usually care about what other people think of us. In some cases, these interpersonal concerns can bias our reasoning, but I think they're also just as likely to encourage us to think about whether or not we can justify ourselves to others. And as other philosophers have noted,[7] this desire to justify ourselves to others is apt to cause us to think critically about our attitudes and actions. So, I see no reason to think that just because we care about our reputation and getting along with others, we won't be inclined to question our attitudes and actions. On the contrary, it's precisely because we have these concerns that we are often inclined to do this.

Of course, as Haidt and others have noted, there's a potential downside to this desire to justify ourselves to others, which is that it can cause us to confabulate when we don't know what the reasons for which we do things are. And insofar as confabulations involve false beliefs about the causes of our attitudes and actions, it might be thought that they're irrational or in some other way inimical to agency. However, I think it's important again to consider the long-term effects

[6] There are, of course, differences in the counterfactual stability of these two motives. For example, if someone who is externally motivated thinks that no one will ever find out about her biases, then she presumably wouldn't be motivated to reflect on and regulate their causes. However, someone who is internally motivated would be under those circumstances.
[7] See, e.g., Velleman (2009).

that confabulations can have. Indeed, as Valerie Tiberius (2013) points out, what was once a confabulated reason can become a motivating reason. So, even if confabulated reasons don't play a role in the production of our attitudes and action, they can still play an important role in their maintenance.

And there's a decent amount of empirical evidence to back this claim up. For example, Wilson and his colleagues (1989) found that people's preferences become more stable and more likely to guide subsequent behavior when they think about reasons for them (whether or not they're the reasons why they have those preferences in the first place),[8] and other studies have shown that the same finding applies to social values like equality, helpfulness, and forgiveness (Maio & Olson, 1998; Maio *et al.,* 2001; Bernard, Maio, & Olson, 2003). As Maio *et al.* (2001) explain, "generating reasons for a value motivates pro-value behavior because individuals become convinced that the value is 'rational' and not just ideological. That is, generating reasons for a value provides concrete examples of why behaving consistently with the value is sensible and justified" (p. 114). So, while confabulations might seem like mere rationalizations at first, they can have significant effects on our agency.[9]

2.3.Haidt's Argument Reconsidered

Before we move on, I want to briefly take stock of what I've tried to argue so far. At a general level, it seems to me that Haidt's argument for thinking that we often aren't willing to engage in the kind of critical reflection that reflectivists associate with human agency is based on some common misconceptions about what exactly reflectivists are committed to. On this

[8] Wilson and his colleagues (2002) also famously found that reflecting on our reasons can also cause us to make worse decisions than we otherwise would've made. I'll return to this finding in the next section, but for now note that this effect is likely to obtain only when the person hasn't put much thought into their choices in the relevant domain.

[9] Tiberius (2013) makes a similar point.

understanding of the view, the idea is that, just in virtue of being rational creatures, we want to know whether or not our attitudes and actions are based on good reasons, and it's this rational concern that makes us inclined to constantly ask ourselves whether we should have the beliefs that we have or whether we should perform the actions that we want to perform. But I think this view of reflective agency is mistaken for two main reasons. First, we shouldn't assume that we're always inclined to question our attitudes and actions. Instead, it seems to me that reflectivists can (and indeed should) think that we often use our reflective capacities selectively. Second, we also shouldn't assume that our motivation to engage in this sort of reflection that reflectivists associate with human agency has to be based on a desire to get things right. Instead, it could be based on familiar interpersonal concerns. So, just because our reasoning is often driven by coherence and relatedness motives, we should conclude that we often aren't willing to question our attitudes and actions.

3. The Problem of Unreliable Reflection

But even if we are often willing to engage in critical reflection, why should we think that it tends to improve our agency? Sometimes the worry here is cashed out in terms of the biological function of reasoning. Thus, for example, Haidt (2001) argues that because moral reasoning didn't evolve to serve an epistemic aim, we can't count on it to help us arrive at true beliefs about what morality requires (for a similar view about reasoning in general, and not just moral reasoning, see Mercier & Sperber (2011)). But although these claims about the evolutionary function of reasoning are interesting, I don't think they're directly relevant to debates about the reliability of reflection. That's because things can often be co-opted to serve functions for which they weren't selected (e.g., our feet presumably didn't evolve to help us run, and yet we often use them for this purpose anyway). So, when we're asking about the reliability

of reflection, I take it that what we really want to know is how often do we use it to improve our agency, regardless of what its biological function may be.

Now, one obvious way that we could try to argue that we often use reflection to improve our agency is to show that human reasoning isn't at biased as research in psychology seems to suggest. In the last section, I suggested that something like this might be true when it comes to research on the my-side bias – that we shouldn't assume that people should always respond to evidence in the same way, irrespective of their prior beliefs. So, on this view, the my-side bias (and confirmation bias more generally) turns out to be more rational than it seems. And if this strategy generalizes to other kinds of studies that purport to show that human reasoning is often susceptible to biases, then maybe we should be more optimistic about the reliability of our reflective capacities (or at the very least we don't have very good reason to be pessimistic about them).

But even if this strategy doesn't work, I think there's another strategy that reflectivists can pursue here, which is to point out that what a lot of this research shows, if it's right, is that individual reasoning tends to be biased. However, other people might often be able to help us to notice and correct these mistakes. So, as long as reflectivists deny that human agency requires individual reflection, the odds that we'll be able to use reflection to improve our agency seem to increase quite a bit.

In what follows, what I want to do is pursue this second strategy. In part, that's because I worry that individual reflection is often biased. But it's also because it seems like most of the time when we engage in reasoning, other people are involved. So, I take it that this approach

captures something important about our social nature. Consider how John Doris and Shaun Nichols (2012) put the point:

> Individualism is abundantly evident in the contemporary cognitive science of morality. Think of the social psychology "vignette" studies that have, in no small measure due to the relative economy of their production, been the paradigm most often pursued by philosophical experimentalists studying moral cognition: sneak into the department office under cover of darkness, Xerox the needed copies of your stimulus material, and you're good to go, near enough for free. Now look, for a moment, as our dutiful participants quietly (if unnervingly quickly) fill out their questionnaires: they're hunched over their desks, looking neither left or right, saying nothing, as they work through the intricacies of some moral problematic.
>
> If the silence isn't deafening, it should be. Instead of this monastic setting, go down to your neighborhood bar, where you might find the regulars discussing some or another moral conundrum: the legitimacy of torture say, or the finer points of marital infidelity. Voices are raised; fingers are pointed. And that's our point. In the wild, fold don't do this stuff quietly; they cajole, they plead, they argue, they abuse. And they do all of this *together*. It's hard to imagine something less like this than our dutiful experimental participants, mutely circling numbers on a Likert scale (pp. 19-20).

Doris and Nichols are focusing on moral reasoning here, but their point seems to apply just as well to reasoning in other domains. Thus, if Doris and Nichols are right, then we might expect human reasoning, whether moral or not, to be at its best when it's done with other people. They refer to this view as "collaborativism," and I think it can help reflectivists avoid worries about the reliability of reflection.

4. Collaborativism

As Doris and Nichols note, there are lots of ways in which reasoning might be thought to depend on other people. For example, other people play a huge role in the development and maintenance of our cognitive faculties, and the transmission of knowledge is an essentially social phenomenon (we're all standing on the shoulders of giants, after all). However, when I talk about

"collaborativism," what I have in mind is something more specific, namely that reasoning is often more likely to result in true beliefs when it takes place in a group setting than when it's done individually. Thus, the basic idea behind collaborativism, as I understand it, is that our reflection will be more reliable when it's done with other people than when we do it in a room by ourselves.

So, is collaborativism true? Does group reasoning tend to outperform individual reasoning? From the outset, let me stress that the answer to this question seems to be highly task dependent. That is, on some tasks, individuals appear to outperform groups, whereas on other tasks groups seems to be better (Thompson, 2008).[10] Consider, for example, work on brainstorming. In these studies, participants are asked to generate as many ideas as they can, to avoid ideas that are vulnerable to criticism, and to try to combine and improve on other people's ideas (Kerr & Tindale, 2004: p.627). Some of the earliest work on brainstorming suggested that groups tend to outperform individuals. But this finding has since been called into question. In particular, it appears that when all of the individual performances are considered in the aggregate (that is, when you compare groups to what are known as "nominal" groups), the nominal groups do a better job. So, in other words, although a group of people brainstorming together is likely outperform any one individual, they're not likely to outperform groups of individuals working separately.

And there are several factors that help to explain why. First, it's obviously harder to free ride when you're asked to perform a task on your own, so nominal groups have a motivational advantage. They also can work on several ideas at once, which means they're less prone to

[10] This has led Thompson to suggest that "the general conclusions of surveys of the empirical research so far is that taken together the findings are mixed or inconclusive." But others are less pessimistic (e.g., Mercier & Landemore, 2013) and it does indeed seem like we can make some predictions about when groups will do better than individuals.

"production blocking" (Diehl & Stroebe, 1987). Finally, nominal groups don't face the various social or emotional factors (e.g., social anxiety and fear of being judge) that inhibit people from contributing in a group setting. So, for all of these reasons, the consensus in psychology now is that if you want a productive group of brainstormers, it's best to ask them to go at it alone (Kerr & Tindale, 2004).

At the same time, most psychologists also seem to agree that groups are better at tasks that admit of obviously right answers. In these situations, the prevailing theme is that "truth wins," meaning that as long as someone in the group knows what the right answer is, is willing and able to explain it with the other members of the group, and those other members have the wherewithal to understand that explanation, they'll accept it (Laughlin & Elis, 1986; Davis, 1969; Moshman & Geil 1998; Sniezek & Henry 1989; Stasser & Dietz-Uhler, 2001). The Wason Selection task is a good example. In this task, participants are asked to test a rule by turning over cards that they think will show whether the rule is true. For example, they might be given the rule "If a card has a number of one side, it has a vowel on the other side," and their job is to test whether every card that has a number on one side also has a vowel on the other. Since a conditional statement is false only if the antecedent is true and the consequent is false, the only way to test this rule is by turning over cards that either don't have a number (that way, if the other side has a vowel, then you know the conditional is false) or that have a vowel (because then the consequent is true and you can find out if the antecedent is true). When individuals perform this task, they're not very good; only about 10% of them get it right. But when groups perform it, that number increases dramatically; about 70% of them get it right ((Moshman & Geil, 1998; Augustinova, 2008; Maciejovsky & Budescu, 2007).

Similar kinds of dramatic differences arise when we compare how individuals and groups perform on brainteasers. For example, in one study, Shaw (1932) asked participants to solve the "cannibals and missionaries" puzzle. The point of this puzzle is to get three missionaries and three cannibals safely from one side of a river to another. But the boat that you have can carry only two of them at a time, and if the missionaries are ever outnumbered by the cannibals, they'll die. Shaw found that while individuals were able to solve the puzzle only about 14% of the time, groups were got the correct answer about 60% of the time (see also Hill, 1982; Stasser & Dietz-Uhler, 2001). Likewise, Schwartz (1995) asked participants to solve four different puzzles that involved spatial reasoning, and he found that groups were 44% more likely to get the answer right. Interestingly, Schwartz also compared the performance of groups of people working together with "nominal" groups (i.e., groups of people working separately), and again he found that the latter were much more likely to solve the puzzles. So, unlike with brainstorming, it looks like groups that are asked to solve brainteasers tend to outperform both individuals and groups of individuals working on their own.

In these cases, what psychologists are testing whether groups are better at tasks that involve theoretical reasoning. But what about tasks that involve practical reasoning? Do groups tend to be better? Again, context matters. For example, when groups of people who already agree on an issue are asked to discuss it, they'll come out of the discussion even more confident in their initial views (Kogan & Wallach 1966; Vinokur & Burnstein, 1978). What's more, during American desegregation, the hope was that diversity in the classrooms would ease racial tensions, but this isn't what happened. Instead, it led to stronger racial in-groups, which only increased the hostility between white and black students (Rogers, Hennigan, Bowman & Miller 1984; Oskamp & Schultz 1998).

But despite these concerns, I think there is some reason to be optimistic about group reasoning in practical domains. For one thing, both of the findings above – i.e., that group discussions lead to polarization and that diversity in the classroom can increase racial tension – are likely to happen only under specific conditions. In particular, polarization is likely to happen only when the members of the group already agree on the issue being discussed. But when they disagree, the discussion is likely to result in depolarization, meaning that they'll come away from discussion less confident in their initial views (Kogan & Wallach 1966; Vinokur & Burnstein, 1978). Likewise, research on diversity in the classroom suggest that the racial tension that resulted from early desegregation efforts was a function of the classroom environment – it was highly competitive. Indeed, studies have shown in more collaborative environments (e.g., when black and white students were asked to work on a project together), racial tension is greatly reduced (Aronson & Bridgeman 1979).

Research using deliberative polls has also shown that group discussions can have a significant impact on people's political beliefs. In these studies, the participants are usually asked about their views on a range of contentious political issues before and after they take part in moderated small group discussions or attend panel discussions with experts who represent a range of perspectives on the topic. In general, what these studies show is that people will often change their minds as a result of these activities. As Akerman and Fishkin (2004) tell us, "it is not unusual for deliberation to significantly change the balance of opinion on two-thirds of the policy questions. And more than half of the respondents typically change their positions on particular policy items after sustained conversations" (p. 52). For example, at one of these events in Austin in 1996, researchers found that, after these discussions, participants were less likely to favor a flat tax; more likely to think that "economic pressure" is "the "biggest problem facing the

American family" and less likely to think that the biggest problem is the "breakdown of traditional values"; and more likely to think that "divorce should be made harder to get" in order to strengthen families (Akerman & Fishkin, 2004; see also Doris & Nichols 2012). Now, it's obviously controversial whether all of these are positive changes, but at the very least what these results show is that group discussions of political issues aren't entirely futile – they can have an impact on people's political beliefs.

Finally, a number of studies have shown that peer discussions can play an important role in moral development (Berkowitz & Gibbs, 1985; Damon & Killen, 1982; Nucci, 1985; Blatt & Kohlberg, 1975; Leman and Duveen, 1999; see also Mercier 2011 for a discussion of this research). In particular, what these studies suggest is that when children are asked to discuss certain moral "dilemmas" in a group, they tend to make different and arguably better moral judgments than they would have made had they been asked to respond to the dilemma on their own. Take, for example, the case of Heinz, who steals medicine that he can't afford because his wife is dying of cancer. When children are asked to discuss Heinz's behavior with their peers, they are more lenient than they would have been had they are asked to evaluate his conduction their own (Leman & Duveen, 1999). They also tend to take it easier on John, who accidentally broke six cups when he was opening the door to the kitchen, than they do on David, who broke six cups while he was trying to steal some cookies. In either case, the shift in judgment that we find between groups and individuals strikes me as a clear improvement. Maybe stealing is always wrong, but Heinz at least has an excuse (if not a justification) for his behavior, and so leniency makes sense. Likewise, John does seem less blameworthy than David is for breaking the cups, and so the fact that children are more likely to track these differences as a result of peer discussion seems to be a good thing (Mercier, 2011).

But what explains these findings? That is, if group dynamics tend to improve our reasoning, why? There are a few related factors. One of them is informational: when we reason with other people, we're likely to be exposed to information that we wouldn't have considered on our own, and it's not implausible to think that our reasoning tends to be better when it's more informed (Doris & Nichols, 2012). Indeed, this appears to be one of the main reasons why deliberative polling works; people get exposed to lots of new information, which in turn causes them to change their beliefs (Luskin et al., 2008).

Similarly, the fact that we're good at coming up with reasons to believe things that we want to believe isn't as problematic in group settings. In fact, as long as the other people are motivated to disagree with us, it can be quite productive. After all, in these cases, because our interlocutors will also be good at coming up with reasons to support their own beliefs, we're likely to be exposed to arguments that we wouldn't have considered on our own – and the same goes for the other people as well. So, although motivation might often bias our reasoning when we're alone, it can serve good epistemic aims in group settings, particularly when group is diverse (Doris & Nichols, 2011).

Finally, other people can also help us to see problems with our reasoning that we wouldn't have otherwise noticed. Again, this is especially likely to occur when those people already disagree with us. That's because, as the studies on my-side bias suggest, our ability to evaluate arguments is asymmetrical: we're pretty good at evaluating arguments that go against our beliefs, but not very good at evaluating arguments that support them. Hence, when we're reasoning with people who don't already agree with us, there's a decent chance that they'll be able to identify any problems with our arguments, and *vice versa*. So, it's not just that diverse groups are able to generate a lot of different arguments for and against the various positions that

the members of the group hold, but that they're in a better position to identify problems with those arguments (Sperber & Mercier, 2011).

All told, then, research suggests that we can expect group reasoning to be more reliable than individual reasoning in settings where a diverse set of view is represented among the members of the group. In these cases, many of the biases that raise concerns about individual reasoning – like our tendency to seek out and recall reasons to believe things that we want to believe – are apt to be corrected. So, given that people often worry about reflectivism precisely on the grounds that we typically use reasoning simply to rationalize our own attitudes and actions, one plausible response that reflectivists can give here is to point out that other people can help us to overcome these problems. Thus, as long as reflectivists are willing to say that exercises of agency needn't require *individual* reflection, but instead might often require us to reason with other people, then worries about motivated reasoning and the reliability of our reflective capacities lose some of their force.

I admit, however, that there are some limits to this defense of reflectivism. First, it could be that we're often reluctant to discuss our views with people who disagree with us, in which case group reasoning won't be very productive – in fact, it could often be worse than individual reasoning, as research on polarization suggests (see again Kogan & Wallach 1966; Vinokur & Burnstein, 1978). Second, even if it's true that group reasoning tends to be better than individual reasoning, you still might not be very impressed by this claim. To be sure, it's perfectly compatible with thinking that group reasoning is fairly unreliable; it's just not as unreliable as individual reasoning is. So, it might be thought that reflectivists need to go one step further. It's not enough to establish the contrastive claim that group reasoning is more reliable than

individual reasoning. They also need to show that it tends to be fairly reliable relative to other kinds of belief forming processes.

Both of these concerns raise important empirical questions (namely, to what extent are we willing to discuss our beliefs with people who disagree with us and is reasoning, whether it's done individually or in groups, more reliable than other kinds of belief-forming processes), and I don't think they can be easily dismissed. Nevertheless, I want to make two points in response. First, I think it's important to note that the debate between reflectivists and their critics has largely focused on the role that individual reflection plays in our everyday lives. In part, that's because most of the studies on which this debate is based have also focused on individual reasoning (Doris & Nichols, 2012). So, the emphasis that we find in the empirical literature gets reflected in the philosophical debate. But I suspect that there's another reason why this debate has focused on individual reflection, which is that it's often assumed that reflectivists are committed to thinking that human agency depends on this kind of reflection. In fact, many critics of reflectivism themselves seem to be sympathetic with the idea that group dynamic can improve our reasoning. Doris (2015) and Mercier and Sperber (2011) are clear examples of this, and I take it that Haidt (2001) can be included in this group as well. So, the main question in this debate hasn't been whether group reasoning can improve our agency (most people involved seem to think it can); it's whether individual reasoning can do the same. Now, maybe it's a mistake to think that group reasoning is especially trustworthy, but even if that's right, it would be a problem for lots of people involved in this debate, not just reflectivists, and settling this issue would seem to take us well outside of its scope.

Second, as Doris and Nichols noted above, this focus on individual reasoning in both the empirical literature and the related philosophical debate obscures the extent to which our

everyday reasoning is a social phenomenon. Usually when we face a practical problem, we don't just sit in a room by ourselves reflecting on what to do. We talk to other people about it – to friends, family members, colleagues, and maybe even strangers. In some cases, we might ask them about a contentious political issue, but this certainly isn't always true. We might instead want to know how to handle a difficult student, how to resolve a conflict between work and our personal lives, or where to get the best coffee in town. Now, if we're surrounded by sycophants, this process probably won't be too reliable– people will tell us only what we want to hear. But any good personal relationship will occasionally require tough love, and most of us are lucky enough to have such relationships.

5. Conclusion

To wrap things up, I think it might help to contrast the sort of reflectivism that I think is empirically defensible with what I take to be a common caricature of the view. On this version of the view, not only does human agency require us to learn about the processes behind our attitudes and actions through introspection, but we are all, just by dint of being rational, willing and able to reflect critically on those processes on our own, without being spurred on or assisted in any other way by other people, if only for the sake of figuring out what we really ought to think, feel, or do. In other words, this view assumes that (i) human agency requires us to have direct access to the objects of reflection, (ii) such reflection is usually done individually, and (iii) it's usually motivated by accuracy.

Like I said, this version of reflectivism is just a caricature of the view, and it's not clear to me that any actual reflectivists accept all of the relevant claims (or at any rate would accept them if pressed). Still, it's not a complete accident that reflectivists are often thought to hold at least

some of these claims. Indeed, the way they talk about agency often gives the impression that this is what they think. To see this, just consider again what Korsgaard (2009) has to say about human agency:

> [H]owever it may be with the other animals, there is no question that we human beings are self-conscious in a very particular way. We are aware, not only *that* we desire or fear certain things, but also that we are inclined to act in certain ways on the *basis* of these desires or fears. We are conscious of the potential grounds of our actions, the principles on which our actions are based, *as potential grounds*. And this, as I have argued elsewhere, sets us a problem that the other animals do not have. For once we are aware that we are inclined to act in a certain way on the ground of a certain incentive, we find ourselves faced with a decision, namely, whether we should do that. We can say to ourselves: "I am inclined to do act-A for the sake of end-E. But should I?" The same applies in the theoretical realm. An intelligent but non-rational animal may be moved to believe or expect one thing when he perceives another, having learned to make a certain causal connection or association between the two things in the past. But he does not think about that principle of association itself, and ask himself whether he should allow it to govern his thinking. But as rational animals we are aware that we are inclined to take one thing as evidence for another, and therefore we can ask whether we should. Our awareness of the workings of the grounds of our beliefs and actions gives us control over the influence of those grounds themselves. (pp. 115-16)

At no point does Korsgaard explicitly commit herself to (i)-(iii), but I think it's quite natural to read her as holding these claims. For example, she talks about our being aware or conscious of the "grounds of our actions," but she doesn't say how such awareness is acquired. Nevertheless, we might take her to be assuming that we usually have direct access to such grounds, and certainly there's nothing in this quotation that suggests that she thinks otherwise. Similarly, she also seems to think as soon as we're aware of the grounds of our actions, we'll often be inclined to reflect on them critically. Again, she doesn't say why we would be so inclined or whether such reflection is done individually or with other people, but I think we could easily read into her remarks the assumption that it's something that we do on our own and for the sake of accuracy.

But then again, if I'm right, Korsgaard doesn't have to accept any of these claims. Instead, she could think that we don't have to have introspective access to the objects of our reflection, that reflection is often driven not by accuracy but by our everyday interpersonal concerns, and that we typically engage in such reflection with other people. That's the brand of reflectivism that I prefer, and I think it's a version of the view that's available even to someone like Korsgaard.

To illustrate the differences between these views, let's consider an example. Suppose that you have a difficult student. His views are obnoxious, and he tends to express them in an especially aggressive way, and you've noticed that his behavior is starting to affect the other students – they seem annoyed or intimidated by him and are less willing to participate in class. So, you face a practical problem — what to do about the student? — and let's just imagine that, because you prefer to avoid conflicts whenever possible, your first instinct is to do nothing. It's still early in the term, you think to yourself, and maybe things will eventually work themselves out. But is that what you should do?

If we take the individualist approach, there are several ways in which you could go wrong. First, you might not know what's motivating you do to nothing, and even if you do, you still might not want to reflect on it critically. But let's suppose that neither of these things is true; as it so happens, you do know that you're inclined to do nothing because of your aversion to conflict and you are sufficiently motivated, if only for the sake of accuracy, to think critically about whether you should act on this inclination. Still, you're not in the clear just yet. After all, whether or not you realize it, your aversion to conflict is likely to bias your reasoning in ways that stack the deck in its favor. For example, because of this aversion, you might find it easier to think of reasons to do nothing, or if you do think of reasons to intervene, they might not seem as

weighty to you. So, even if you make a good faith attempt to reflect on whether you should something about this student, you might decide that you shouldn't because that's what you wanted to do all along.

However, if we take the approach to reflective agency that I prefer, then it's plausible to think that many of these problems are at least less likely to occur. First, you won't have to rely on introspection alone to figure out why you're inclined to do nothing, and your motivation to reflect on your aversion to conflict, should you learn about it, doesn't have to stem from accuracy goals. It could instead come from any number of sources (e.g., maybe you're simply worried about your teaching evaluations or you want to encourage the other students to participate more so that you don't have to lecture as much). At any rate, while there's no guarantee that you'll know why you want to do nothing or that you'll be inclined to think critically about this desire if you are aware of it, it still seems safe to say that both of these things are more likely to occur on the model that I prefer. What's more, as long as we assume that you don't have to (and indeed shouldn't) reflect on what you should do individually, you will also be more likely to overcome many of the biases that are likely to interfere with your reasoning otherwise. For example, suppose that rather than thinking about how to handle this situation on your own, you ask your colleagues for advice. Maybe one of them has had similar experiences and can tell you first-hand that doing nothing only made the situation worse, or maybe, since they don't share your motivation, they'll at least be able to raise considerations that you hadn't thought of or were quick to dismiss. Whatever the case, as long as you're willing and able to ask other people for advice, there's a decent chance that your reflection be more reliable as a result, and that's simply because they will probably be in a better position to notice and correct any biases that would interfere with your reasoning were you to go at it alone.

Chapter 6

Reflectivism Revisited

In this dissertation, I have defended a version of reflectivism against recent challenges that emerge from research in social psychology. In this final chapter, what I want to do is review some of the main conclusions that I drew in each chapter and present a more comprehensive account of human agency.

In Chapter One, I sketched what I take to be the three most prominent challenges that recent research in social psychology presents to reflectivists views of human agency. The first challenge, which I called the Problem of Automaticity, is based on the finding that a lot of what we think, feel, and do issues from automatic processes. If this is true, then you might wonder how often we do things for reasons. Can our attitudes and actions issue from automatic processes and still be had or done for reasons, and if so, under what conditions does this happen and how often do those conditions obtain in our everyday lives?

The second challenge that I focused on, which I called the Problem of Self-Knowledge, concerns our lack of self-knowledge. In particular, one of the main findings that comes out of research on automaticity is that people often aren't in a good position to report accurately on the causes of their attitudes and actions. But if that's right, then you might wonder to what extent are we really able to use reflection to improve our agency. After all, doing so would seem to require us to have some awareness of the causes of our attitudes and actions, but that's precisely the sort of self-knowledge that we seem to lack. So, insofar as reflectivists are committed to the claim that we can use reflection to monitor and control the influences on our attitudes and action, they seem to be mistaken.

The final challenge that I focused on, which I called the Problem of Unreliable Reflection, extends these concerns about the limits of self-reflection. More specifically, even if we assume that we do often know what the causes of our attitudes and actions are, there's still no guarantee that we'll use reflection to improve our agency. And that's because research suggests that human reasoning is subject to a wide range of biases. As a result, it looks like we often don't want to think critically about our own attitudes and actions, and even when we do, reflection might still often lead us astray.

In Chapter Two, I focused on recent attempts to resolve the Problem of Automaticity while still holding on to the claim that our ability to do things for reasons depends on our reflectivists capacities. In general, what all of these approaches have in common is the idea that there is more than one way in which our ability to do things for reasons can depend on our reflective capacities. Critics of reflectivism often assume that the relationship has to be direct – that whenever we do something for a reason, some kind of deliberation or reflection has to be the proximate cause of our behavior. But this is just one among many options, and once we recognize that the relationship between reflection and our ability to do things for reasons can be indirect, the fact that our attitude sand actions issues from automatic processes itself isn't a threat to the claim that we often do things for reasons.

However, I argued that, despite their advantages over views that claim that we have to engage in reflection whenever we do something for a reason, there are good reasons to reject these views. The first worry is primarily conceptual: since reflection is not only something that we do, but indeed something that we typically do for reasons, we need some account of how these acts of reflection themselves can be done for reasons. And yet if we accept a version of the reflection view, the only answer we can give is that those acts of reflection stand in the right

relationship to yet further acts of reflection. But because those further acts of reflection will often be done for reasons as well, this response only pushes the problem back a step, for now we need to know in virtue of what are those even further acts of reflection done for reasons, and so on, *ad infinitum*. So, as long as we accept that our ability to do things for reasons depends on our reflective capacities and that reflection is often something that we do for reasons, we're going to run into some kind of regress problem.

However, even if advocates of the reflection view can either block or learn to live with the regress, I tried to show that there are still good, empirical reasons to be skeptical of it – or rather there are good, empirical reasons to be skeptical of at least two prominent versions of this view. First, I argued that we can't account for the full range of behavior that is automatic-and-yet-still-done-for-a-reason by appealing to the role that our reflective capacities play in the acquisition and development of these process, as advocates of the Development View try to do, because some of the things that we do for reasons are guided by psychological mechanisms that we've acquired or developed through implicit learning processes (that is, learning processes that didn't involve any kind of reflection). As a result, it seems like the Development View is underinclusive: there are times when we do things for reasons, and yet our behavior don't satisfy the conditions that this view lays out.

One thing that this argument might suggest is that we shouldn't try to locate our ability to do things for reasons in actual acts of reflection at all, whether they're proximate or distal. Instead, perhaps what we should say is that it simply depends on our satisfying a counterfactual condition: that our ability to do things for reasons requires us to act in response to considerations that we would recognize as reasons if we were to reflect on them. Although this sort of view is less psychologically demanding than any version of the reflection view that requires us to engage

in actual reflection, I still don't think we should accept it. That's because the reasons for which we do things aren't always good reasons, even by our own lights. Thus, there will be times when we do things for reasons, but we wouldn't recognize the considerations on which we acted as reasons for our behavior. So, again, it seems like this version of the reflection view will be underinclusive as well.

What all of these problems with the reflection view suggest to me is that the best way to respond to the Problem of Automaticity isn't to try to show that automatic processes often stand in the right relationship to our reflective capacities and can thereby cause us to do things for reasons. Rather, it's to try to show that automatic processes themselves can guide rational thoughts and actions, independently of their relationship to our reflective capacities. In Chapter Three, I defend this sort of view by developing ideas found in recent work by Peter Railton and other philosophers. In particular, I argued for two main claims: (1) that recent developments in research on reinforcement learning and expert decision-making suggest that many of our attitudes and actions are guided by sophisticated, model-based learning processes that select actions based on their expected utility and (2) that when we act on the basis of these processes, we can be said to do things because we recognize reasons to do them (indeed, because we *believe* that we have reasons to do them). So, according to the view that I defend, our ability to do things for reasons doesn't depend on our reflective capacities. It simply requires us to do things in response to our recognition of reasons. And since I think this is often true of us when we act on the basis of automatic processes, we can reconcile research on automaticity with the idea that we often do things for reasons.

But I admit that model-based learning systems are only going to be as good as the environment in which they're trained. Thus, given that we live in a world that involves unjust

social structures, we can expect them to have certain biases. As a result, these systems won't always cause us to do things for good reasons, and so it would be nice to have some way to monitor and control their influences on us. Traditionally, this is the role that reflection has been thought to play in our practical lives, and I think it's a role that it's well-suited to play. However, in order to defend this view, I think we need to revise some common assumptions about what reflection is and how it works. So, in the second half of this dissertation, I try to identify and motivate those revisions.

In Chapter Four, I focus on the Problem of Self-Knowledge. In the first instance, the problem here is that we often seem to lack access to the processes behind our attitudes and actions, and since we can't use reflection to improve our agency unless we have access to those process, it seems like we shouldn't expect reflection to play a regulative role in our everyday lives. However, in response to this challenge, I argued that most psychologists, including Nisbett and Wilson, don't seem to accept that we typically lack access to the processes behind our attitudes and actions. Instead, what they seem to accept is a more restricted claim: namely, that we often lack direct access to those processes. But that still leaves open the possibility that we often learn about the causes of our attitudes and actions by using indirect methods, and I argued that research on the self-regulation of implicit biases gives us reason to be optimistic about the scope of our indirect self-knowledge.

However, even if we do have a lot of indirect knowledge of the causes of our attitudes and actions, it might still be argued that this doesn't show that we can use reflection to improve our agency, because, according to this response, reflection requires us to have direct access to its objects. But I argued that this view, though perhaps widely held, is nevertheless incompatible with two other common assumptions about reflection and its objects: first, that we can reflect on

our character traits and other behavioral tendencies and, second, that we don't learn about our character traits and other behavioral tendencies directly. So, it seems like one of these assumptions has to go, and since it seems to me that most reflectivists would want to accept the latter two, I take it that they would therefore have to reject the claim that we have to have direct access to the objects of reflection. Thus, given that I think we have a decent amount of indirect knowledge of the causes of our attitudes and actions and that reflection doesn't require us to have direct access to its objects, I argued that we will often be in a good position to use reflection to monitor and control their influences on us.

But we don't always want to use reflection to think critically about the causes of our attitudes and actions, and even if we do, because reflection is often biased, there's no guarantee that we'll always get things right. So, even if we're in a good position to use reflection to improve our agency, why think we will? In Chapter Five, I argued that we can at least mitigate these worries by focusing on the social aspects of reflective agency. In particular, because other people are often in a better position to notice our biases than we are, especially if they're motivated to defend a different view, our reasoning will often be better when it's done in group settings. As a result, I think reflectivists would do well to shift away from a model of reflective agency that conceives of it as an individual achievement and to focus instead on the ways in its can be improved by group dynamics.

If all of this is right, then the version of reflectivism that we end up with is one that is in several respects quite different from the view that people often associate with reflectivism. For one thing, I don't think that reflectivists have to be committed to the claim that our ability to do things for reasons requires us to engage in reflection. Instead, they can say that it depends on the parts of our psychology that we have in common with lots of different animals. Of course, one

consequence of this view is that reflectivists can't also say that our ability to do things for reasons is what makes human agency special. But that's not much of a cost – surely our furry and feathered friends can do things for reasons too! For another thing, I don't think reflectivism should be confused with the claim that human agency is at its best when we're sitting in a room by ourselves, carefully thinking about the causes of our attitudes and actions via introspection if for no other reason than to make sure that they're reasonable. As cliché as it is, we're social creatures, and reflectivists certainly needn't think otherwise. Indeed, if I'm right, not only do we often learn about the objects of our reflection thanks to other people, but they can also correct many of the biases that affect individual reasoning.

Works Cited

Ackerman, B. & Fiskin, J. S., 2004., *Deliberation Day*. Yale University Press.

Augustinova, M. 2008, "Falsification Cueing in Collective Reasoning: Example of the Wason Selection Task," *European Journal of Social Psychology*, 38(5), 770-85.

Anderson, E., 1995, "Feminist Epistemology: An Interpretation and a Defense," *Hypatia*, 10(3), 50-85.

Anscombe, E., 1957, *Intention,* Harvard University Press.

Arpaly, N., 2003, *Unprincipled Virtue: An Inquiry into Moral Agency*, Oxford University Press.

Arpaly, N., & Schroeder, T., 2012, "Deliberation and Acting for Reasons," *The Philosophical Review*, Vol. 121, No. 2, 209-239.

Aronson, E., & Bridgeman, D., 1979, "Jigsaw Groups and the Desegregated Classroom: In Pursuit of Common Goals," *Personality and Social Psychology Bulletin*, 5(4), 438-46.

Arpaly, N., & Schroeder, T., 2015, *In Praise of Desire*, Oxford University Press.

Bailenson, J. N., & Yee, N., 2005, "Digital Chameleons: Automatic Assimilation of Nonverbal Gestures in Immersive Virtual Environments," *Psychological Science*, 16(10), 814-19.

Balleine, B. W., & Dickinson, A., 1998, "Goal-Directed Instrumental Action: Contingency and Incentive Learning and Their Cortical Substrates," *Neurophamacology*, 37(4-5), 407-19.

Bargh, J., 1994, "The Four Horsemen of Automaticity: Awareness, Intention, Efficiency, and Control in Social Cognition," in R. S. Wayer, Jr. & T. K. Srull (eds.), *Handbook of Social Cognition: Basic Processes; Applications*, 1-40.

Bargh, J. & Chartrand, T., 1999, "The Unbearable Automaticity of Being," *American Psychology,* 54(7), 462-79.

Barry, M., 2007, "Realism, Rational Action, and the Humean Theory of Motivation," *Ethical Theory and Moral Practice*, 10(3), 231-42.

Bateson, M., Nettle, D., & Roberts, G., 2006, "Cues of Being Watched Enhance Cooperation in a Real-World Setting," *Biology Letters*, 2, 412-14.

Beach, L. R., & Mitchell, T. R., 1987, "Image Theory: Principles, Goals, and Plans in Decision Making," *Acta Psychologica*, 66(3), 201-20.

Bechara, A., Damasio, H., & Damasio, A. R., 1997, "Deciding Advantageously Before Knowing the Advantageous Strategy," *Science*, 275(5304), 1293-95.

Berkowitz, M. W., & Gibbs, J. C., 1985, "The Process of Moral Conflict Resolution and Moral Development," *New Directions for Child Development*, 29, 71-84.

Bernard, M. M., Maio, G. R., & Olson, J. M., 2003, "Effects of Introspection About Reasons for Values: Extending Research on Values-as-Truisms," *Social Cognition*, 21(1), 1-25.

Bertrand, M. & Mallainathan, S., 2004, "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experience on Labor Market Discrimination," *American Economic Review*, 94(4), 991-1013.

Blatt M. M., & Kohlberg L., 1975, "The Effects of Classroom Moral Discussion upon Children's Level of Moral Judgment," *Journal of Moral Education*, 4(2),129–61.

Bornstein, R. F., 1989, "Exposure and Affect: Overview and Meta-Analysis of Research, 1968-1987," *Psychological Bulletin*, 106(2), 265-89.

Bornstein, R. F., & D'Agostino, P. R., 1992, "Stimulus Recognition and the Mere Exposure Effect," *Journal of Personality and Social Psychology*, 63(4), 545-52.

Brink, D., 2003, *Perfectionism and the Common Good: Themes in the Philosophy of T.H. Green*, Clarendon Press

Brinol, P., & Petty, R. E., 2009, "Changing Attitudes on Implicit Versus Explicit Measures: What is the Difference?" In R. E. Petty, R. H. Fazio, P. Brinol (eds.), *Attitudes: Insights from the New Implicit Measures*, Psychology Press, 285-326.

Brownstein, M., 2014, "Rationalizing Flow: Agency in Skilled Unreflective Action," *Philosophical Studies,* 168(2), 545-68.

Brownstein, M., 2016, "Implicit Attitudes, Social Learning, and Moral Credibility," In J. Kiverstein (ed.), *The Routledge Handbook of Philosophy of the Social Mind*, Routledge, 314-35.

Burns, M. D., Monteith, M. J., & Parker, L. R., 2017, "Training Away Bias: The Differential Effects of Counterstereotype Training and Self-Regulation on Stereotype Activation and Application," *Journal of Experimental Psychology*, 73, 97-110.

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K., 2012, "Sequential Priming Measures of Implicit Social Cognition: A Meta-Analysis of Associations with Behavior and Explicit Attitudes." *Personality and Social Psychology* Review, 16, 330-50.

Chang, R., 2002, "The Possibility of Parity," *Ethics*, 112(4), 659-688.

Chartrand, T. L., & Bargh, J. A., 1999, "The Chameleon Effect: The Perception–Behavior Link and Social Interaction," *Journal of Personality and Social Psychology,* 76(6), 893–910.

Chartrand, T. L., & Larkin, J. L., 2013, "The Antecedents and Consequences of Human Behavioral Mimicry," *Annual Review of Psychology*, 64, 285-308.

Chen, S.., Shechter, D., & Chaiken, S., 1996, "Getting at the Truth or Getting Along: Accuracy-Versus-Impression-Motivation Heurist and Systematic Processing," *Journal of Personality and Social Psychology*, 71(2), 262-75.

Churchland, P. S., & Suhler, C. L., 2014, "Agency and Control: The Subcortical Role in Good Decisions," In W. Sinnott-Armstrong (ed.), *Moral Psychology, Vol. 4:L Free Will and Moral Responsibility*, MIT Press, 309-26.

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., 2007, "The Influence of Stereotypes on Decisions to Shoot," *European Journal of Social Psychology*, 37, 1102-17.

Crockett, M. J., 2013, "Models of Morality," *Trends in Cognitive Sciences*, 17(8), 363-66.

Cushman, F., 2013, "Action, Outcome, and Value: A Dual-System Framework for Morality," *Personality and Social Psychology Review*, 17(3), 273-92.

Czopp, A. M., Monteith, M. J., & Mark, A. Y., 2006, "Standing Up For a Change: Reducing Bias Through Interpersonal Confrontations," *Journal of Personality and Social Psychology*, 90(5), 784-803.

Damon, W., & Killen, M., 1982, "Peer Interaction and the Process of Change in Children's Moral Reasoning," *Merrill-Palmer Quarterly,* 28(3), 347-67.

Dancy, J, 2000, *Practical Reality*, Oxford University Press.

Davidson, D. 1963, "Actions, Reasons, and Causes," *The Journal of Philosophy*, 60(23), 685-700.

Davidson, D., 1973, "Freedom to Act", in T. Honderich (ed.), *Essays on Freedom of Action*, Routledge, 139-56.

Davis, J. H., 1969, *Group Performance*, Addison-Wesley.

DeCoster, J. & Claypool, H. M., 2004, "A Meta-Analysis of Priming Effects on Impression Formation Supporting a General Model of Informational Biases." *Personality and Social Psychology Review*, 8, 2-27.

Diehl, M. & Stroebe, W., 1987. "Productivity Loss in Brainstorming Groups: Toward the Solution of a Riddle." *Journal of Personality and Social Psychology*, 53, 497-509.

Dienes, Z., & Altman, G., 1997, "Transfer of Implicit Knowledge Across Domains: How Implicit and How Abstract?" in D. Berry (ed) *How Implicit is Implicit Learning?* Oxford University Press, 1-32.

Dijksterhuis, A., 2004, "Think Different: The Merits of Unconscious Thought in Preference Development and Decision Making," *Journal of Personality and Social Psychology*, 87(5), 586-98.

Doll, B. B., Simon, D. A., & Daw, N. D., 2013, "The Ubiquity of Model-Based Reinforcement Learning," *Current Opinion in Neurobiology*, 22(6), 1075-81.

Doosje, B. Spears, R., & Koomen, W., 1995, "When Bad Isn't All Bad: Strategic Use of Sample Information in generalization and Stereotyping," *Journal of Personality and Social Psychology*, 75, 568-84.

Doris, J. M., 2009, "Skepticism about Persons," *Philosophical Issues,* 19, 57-91.

Doris, J. M., 2015, *Talking to Our Selves*, Oxford University Press.

Doris, J. M., & Nichols, S. B., 2012, "Broad-Minded: Sociality and the Cognitive Science of Morality," In E. Margolis, R. Samuels, S. Stich (eds.), *The Oxford Handbook of Philosophy of Cognitive Science*, Oxford University Press, 425-53.

Dretske, F., 1988, *Explaining Behavior: Reasons in a World of Causes*, MIT Press.

Dreyfus, H. L., 2005, "Overcoming the Myth of the Mental: How Philosophers can Profit from the Phenomenology of Everyday Expertise," *Proceedings and Addresses of the American Philosophical Association*, 79(2), 47-65.

Egan, A., 2005, "I Can't Believe I'm Stupid," *Philosophical Perspectives*, 19(1), 77-93.

Egan, A., 2008, "Seeing and Believing: Perception, Belief Formation, and the Divided Mind," *Philosophical Studies*, 140(1), 47-63.

Egan, A., 2011, "Comments on Gendler's, 'The Epistemic Costs of Implicit Bias'," *Philosophical Studies*, 156(1), 65.

Evans, J., & Frankish, K., 2009, *In Two Minds: Duel Processes and Beyond*, Oxford University Press.

Fiorillo, C. D., Tobler, P. N., & Schultz, W., 2003, "Discrete Coding of Reward Probability and Uncertainty by Dopamine Neurons," *Science*, 299(5614), 1898-1902.

Fischer, J. M., & Ravizza, M., 1998, *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge University Press.

Fischoff, B., 1977, "Perceived Informativeness of Facts," *Journal of Experimental Psychology: Human Perception and Performance*, 3, 349-58.

Frank, R. H., Gilovich, T., & Regan, D. T., 1993, "The Evolution of One-Shot Cooperation: An Experiment," *Ethology & Sociobiology*, 14(4), 247-56.

Freund, T., Kruglanski, A. W., & Shpitzajzen, A., 1985, "The Freezing and Unfreezing of Impressional Primacy: Effects on the Need for Structure and the Fear of Invalidity," *Personality and Social Psychology Bulletin*, 11, 479-87.

Gardner, W. L., Gabriel, S., & Lee, A. Y., 1999, "'I' Value Freedom, but 'We' Value Relationships: Self-Construal Priming Mirrors Cultural Differences in Judgment," *Psychology Science*, 10(4), 321-26.

Gendler, T. S., 2008a, "Alief and Belief," *Journal of Philosophy*, 105(10), 634-63.

Gendler, T. S., 2008b, "Alief in Action (and Reaction)," *Mind & Language* 23(5), 555-85.

Gendler, T. S., 2011, "On the Epistemic Costs of Implicit Bias," *Philosophical Studies*, 156(1), 33-63

Gigerenzer, G., 1991, "How to Make Cognitive Illusions Disappear: Beyond 'Heuristics and Biases," In W. Stroebe & M. Hewston (eds.), *European Review of Social Psychology*, 2, 83-115.

Gilbert, D. T., 1991, "How Mental Systems Believe," *American Psychologist*, 46(2), 107-19.

Gilber, D. T., Tafarodi, R. W., & Malone, P. S., 1993, "You Can't Not Believe Everything You Read," *Journal of Personality and Social Psychology*, 65(2), 221-33.

Greene, J. D., 2017, "The Rat-a-gorical Imperative: Moral Intuition and the Limits of Affective Learning," *Cognition*, http://dx. doi.org/10.1016/j.cognition.2017.03.004

Gupta, A. S., van der Meer, M. A., Touretzky, D. S., & Redish, A. D., 2010, "Hippocampal Replay is not a Simple Function of Experience," *Neuron*, 65(5), 695-705.

Haidt, J., 2001, "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment," *Psychological Review*, 108(4), 814-34.

Haley, K. J., & Fessler, D. M. T., 2005, "Nobody's Watching? Subtle Cues Affect Generosity in an Anonymous Economic Game." *Evolution and Human Behavior*, 26, 245–256.

Hill, G.,1982, "Group Versus Individual Performance: Are N + 1 Heads Better Than One? *Psychological Bulletin* 91, 517-39.

Huebner, B., 2009, "Trouble with Stereotypes for Spinozan Minds," *Philosophy of the Social Sciences*, 39(1), 63-92.

Huebner, B., 2016, "Implicit Bias, Reinforcement Learning, and Scaffolded Moral Cognition," in M. Brownstein and J. Saul (eds), *Implicit Bias and Philosophy, Vol. 1: Metaphysics and Epistemology*, Oxford University Press.

Hursthouse, R., 1991, "Arational Actions," *Journal of Philosophy*, 88(2), 57-68.

Ji, D., & Wilson, M. A., 2007, "Coordinated Memory Replay in the Visual Cortex and Hippocampus During Sleep," *Nature Neuroscience*, 10, 100-07.

Kahneman, D., 2013, *Thinking Fast and Slow*, Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A., 1962, "Subjective Probability: A Judgement of Representativeness," *Cognitive Psychology*, 3(3), 430-454.

Kane, R., 1998, *The Significance of Free Will*, Oxford University Press.

Kerr, N. L., & Tindale, R. S., 2004, "Group Performance and Decision Making," *Annual Review of Psychology,* 55, 623-655.

Kitayama, S., & Karasawa, M., 1997, "Implicit Self-Esteem in Japan: Name Letters and Birthday Numbers," *Personality and Social Psychology Bulletin*, 23(7), 736-42.

Klayman, J., & Ha, Y., 1987, "Confirmation, Disconfirmation, and Information in Hypothesis Testing," *Psychological Review,* 94(2), 211-28.

Kruglanski, A. W., & Freund, T., 1983, "The Freezing and Unfreezing of Lay-Inferences: Effects on Impressional Primacy, Ethnic Stereotyping, and Numerical Anchoring," *Journal of Experimental Social Psychology, 19*(5), 448–68.

Kogan, N., & Wallach, M. A., 1966, "Modification of a Judgmental Style Through Group Interaction. *Journal of Personality and Social Psychology,* 4(2), 165-74.

Kornblith, H., 2012, *On Reflection*, Oxford University Press.

Korsgaard, C. 1996, *Sources of Normativity*, Cambridge University Press.

Kunda, Ziva, 1999, *Social Cognition: Making Sense of People*, MIT Press.

Kunst-Wilson, W., & Zajonc, R., 1980, "Affective Discrimination of Stimuli that Cannot be Recognized," *Science*, Vol. 207, No. 4430, 557-58.

Kurth, C., 2018, "Emotion, Deliberation, and the Skill Model of Virtuous Agency," *Mind*, Vol. 33, No. 3, 299-317.

Landy, J. F., & Goodwin, G. P., 2015, "Does Incidental Disgust Amplify Moral Judgment? A Meta-Analytic Review of Experimental Evidence," *Perspectives of Psychological Science*, 10(4), 518-36.

Laughlin, P. R., & Ellis, A. L.. 1986. "Demonstrability and Social Combination Processes on Mathematical Intellective Tasks," *Journal of Experimental Social Psychology,* 22, 177–89.

Leman, P. J., & Duveen, G., 1999, "Representations of Authority and Children's Moral Reasoning," *European Journal Social Psycholugy*, 29(5–6), 557–75.

Levin, I. P., & Gaeth, G. J., 1988, "How Consumers Are Affected by the Framing of Attribute Information Before and After Consuming the Product," *Journal of Consumer Research*, 15(3), 374-78.

Levy, N., 2015, "Neither Fish Nor Fowl: Implicit Attitudes as Patchy Endorsements," *Nous*, 49(4), 800-23.

Levy, N., & Bayne, T., 2004, "Doing Without Deliberation: Automatism, Automaticity, and Moral Accountability," *International Review of Psychology*, 16(3), 209-15.

Lord, C. G., Ross, L., & Lepper, M. R., 1979, "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence," *Journal of Personality and Social Psychology*, 37(11), 2098-2109.

Luskin, R. C., Crow, D. B., Fiskin, J. S., Guild, W. & Thomas, D., 2008, "Report on the Deliberative Poll on 'Vermont's Energy Future'." Center for Deliberative Opinion Research, University of Texas at Austin.

Maciejovsky, B., & Budescu, D. V., 2007, "Collective Induction Without Cooperation? Learning and Knowledge Transfer in Cooperative Groups and Competitive Auctions," *Journal of Personality and Social Psychology, 92*(5), 854–70.

Mandelbaum, E., 2015, "Attitude, Inference, Association: On the Propositional Structure of Implicit Bias," *Nous*, 50(3), 629-658.

Maxwell, S., Lau, M. & Howard, G., 2015, "Is Psychology Suffering from a Replication Crisis? What Does 'Failure to Replicate' Really Mean?" *American Psychologist,* 70(6), 487-98.

Maio, G. R., Olson, J. M., Allen, L., & Bernard, M. M., 2001, "Addressing Discrepancies Between Values and Behaviour: The Motivating Effect of Reasons." *Journal of Experimental Social Psychology*, 37, 104– 117.

Maio, G. R., & Olson, J. M., 1998, "Values as Truisms: Evidence and Implications." *Journal of Personality and Social Psychology*, 74(2), 294– 311.

McClure, S. M., Berns, G. S., & Montague, P. R., 2003, "Temporal Prediction Errors in a Passive Learning Task Activate Human Stratum," *Neuron*, 38(2), 339-46.

McKenzie, C. R. M., & Mikkelsen, L. A., "The Psychological Side of Hempel's Paradox of Confirmation," *Psychonomic Bulletin & Review*, 7(2), 360-66.

Mele, A., 1987, "Intentional Action and Wayward Causal Chains: The Problem of Tertiary Waywardness." *Philosophical Studies*, 51(1), 55-60

Mele, A., 2009, "Mental Actions: A Case Study," in L. O'Brien & M. Soteriou (eds.) *Mental Actions*, Oxford University Press*,* 16-38.

Mele, A. 2014, *Free: Why Science Hasn't Disproven Free Will*, Oxford University Press.

Mercier, H., 2011, "What Good is Moral Reasoning?" *Mind & Society*, 10(2), 131-48.

Mercier, H., & Landemore, H., 2012, "Reasoning is for Arguing: Understanding the Success and Failures of Deliberation," *Political Psychology*, 33(2), 243-58.

Mercier, H., & Sperber, D., 2011, "Why Do Humans Reason? Arguments for an Argumentative Theory," *Behavioral and Brain Sciences*, 34, 57-111.

Miller, D. T., Downs, J. S., & Prentice, D. A., 1998, "Minimal Conditions for the Creation of a Unit Relationship: The Social Bond Between Birthdaymates," *European Journal of Social Psychology*, 28(3), 475-481.

Monteith, M. J., Mark, A. Y., & Ashburn-Nardo, L., 2010, "The Self-Regulation of Prejudice: Toward Understanding Its Lived Character," *Psychological Science*, 18, 524-30.

Monteith, M. J., & Voils, C. I., 1998, "Proneness to Prejudiced Responses: Toward Understanding the Authenticity of Self-Reported Discrepancies," *Journal of Personality and Social Psychology*, 75(4), 901-16.

Moshman, D., & Geil, M., 1998, "Collaborative Reasoning: Evidence for Collective Rationality." *Thinking and Reasoning, 4*(3), 231-48.

Nagel, T. 1979, *The Possibility of Altruism*, Princeton University Press.

Nahmias, E., 2007, "Autonomous Agency and the Threat of Social Psychology," in M. Marraffa, M. Caro, & F. Ferretti (eds.), *Cartographies of the Mind: Philosophy and Psychology in Intersection*, 169-88.

Nahmias, E., 2014, "Is Free Will an Illusion? Confronting Challenges from the Modern Mind Sciences," *Moral Psychology, vol. 4*, Walter Sinnott-Armstrong (ed.), MIT Press, 1-56.

Nickerson, R. S., 1998, "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology*, 2(2), 175-220.

Nisbett, R. E., & Wilson, T. D., 1977, "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review*, 84(3), 231-59.

Nucci, L. 1985, "Future Directions in Research on Children's Moral Reasoning and Moral Education," *Elementary School Guidance & Counseling*, 19(4), 272-82.

Oaksford, M., & Chater, N., 1994, "A Rational Analysis of the Selection Task as Optimal Data Selection," *Psychological Review, 101*(4), 608–31.

Oskamp, S. & Schultz, P., 1998, *Applied Social Psychology,* Prentice Hall.

Payne, B. K., 2001, "Prejudice and Perception: The Role of Automatic and Controlled Processes in Misperceiving a Weapon," *Journal of Personality and Social Psychology*, 81(2), 181-192.

Payne, J. W., Bettman, J. R., & Johnson, E. J., 1993, "The Use of Multiple Strategies in Judgement and Choice," In N.J. Castellan, Jr. (ed.), *Individual and Group Decision Making: Current Issues*, 19-39.

Pelham, B. W., Mirenberg, M. C., Jones, J. T., 2002, "Why Susie Sells Seashells by the Seashore: Implicit Egotism and Major Life Decisions," *Journal of Personality and Social Psychology*, 82(4), 469-87.

Quinn, W., 1994, "Putting Rationality in its Place," *Morality and Action*, Cambridge University Press, 228-255.

Railton, P., 2009, "Practical Competence and Fluent Agency," *Reasons for Action*, David Sobel & Steven Wall (eds.), Cambridge University Press, 81-115.

Railton, P., 2014, "The Affective Dog and its Rational Tale: Intuition and Attunement," *Ethics*, Vol. 124, No. 4, 813-50.

Railton, P., 2016, "Moral Learning: Why Learning? Why Moral? And Why Now?" *Cognition*, http://dx.doi.org/10.1016/j. cognition.2016.08.015

Railton, P., (m.s.), *Toward a Unified Account of Rationality in Belief, Desire, and Action*.

Raz, J., 2002, *Engaging Reason*: *On the Theory of Value and Action*, Oxford University Press.

Reber, A. S., 1989, "Implicit Learning and Tacit Knowledge," *Journal of Experiment Psychology: General*, 219-35.

Reed, L. I., Zeglen, K. N., & Schmidt, K. L., 2012, "Facial Expressions as Honest Signals of Cooperative Intent in a One-Shot Anonymous Prisoner's Dilemma Game," *Evolution and Human Behavior*, 33(3), 200-09.

Rogers, M., Hennigan, K., Bowman, C. & Miller, N., 1984, "Inter-Group Acceptance in Classrooms and Playground settings." In N. Miller and M. Brewer (eds.), *Groups in Contact: The Psychology of Desegregation*. Academic Press.

Sanitioso, R., Ziva, K., & Fong, G. T., 1990, "Motivated Recruitment of Autobiographical Memories," *Journal of Personality and Social Psychology*, 59(2), 229-41.

Sauer, H., 2012, "Educated Intuitions: Automaticity and Rational in Moral Judgment," *Philosophical Explorations*, 15(3), 255-75.

Scanlon, T., 2000, *What We Owe To Each Other*, Harvard University Press.

Scheffler, S., 1994, *Human Morality*, Oxford University Press.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H., 2008, "Disgust as Embodied Moral Judgment," *Personality Social Psychology Bulletin*, 34(8), 1096-1109.

Schultz, W., Dayan, P. & Montague, P. R., 1997, "A Neural Substrate of Prediction and Reward," *Science*, 275(5306), 1593-99.

Schultz, W., & Dickinson, A., 2000, "Neuronal Coding of Prediction Errors," *Annual Review of Neuroscience*, 23, 473-500.

Schwartz, D. L., 1995, "Reasoning About the Reference of a Picture Versus Reasoning about the Picture as a Referent: An Effect of Visual Realism," *Memory & Cognition*, 23(6), 709-22.

Schwitzgebel, E., 2002, "A Phenomenal, Dispositional Account of Belief," *Nous*, 36, 249-75.

Schwitzgebel, E., "Introspection," *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/introspection/

Sechrist, G. B., & Stangor, C., 2001, "Perceived Consensus Influences Intergroup Behavior and Stereotype Accessibility," *Journal of Personality and Social Psychology*, 80(4), 645-54.

Sehon, S., 2017, *Free Will and Action Explanation*, Oxford University Press.

Seigel, S., 2017, *The Rationality of Perception*, Oxford University Press.

Shaw, M.E., 1932, "A Comparison of Individuals and Small Groups in the Rational Solution of Complex Problems," *The American Journal of Psychology*, 44, 491-504.

Sher, S., & McKenzie, C. M., 2008, "Framing Effects and Rationality," In N. Chater & M. Oaksford (eds.), *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*, Oxford University Press, 79-96.

Smith, M., 1994, *The Moral Problem*, Blackwell Publishing.

Smith, M., 1998, "The Possibility of Philosophy of Action," in J. Bransen & S. Cuypers (eds.), *Human Action, Deliberation and Causation*, Kluwer Academic Publishers, 17-41.

Sniezek, J. A., & Henry, R. A., 1989, "Accuracy and Confidence in Group Judgment<" *Organizational behavior and human decision processes,* 43(1), 1-28.

Stasser, G., & Dietz-Uhler, B., 2001, "Collective Choice, Judgment, and Problem Solving," In M. A. Hogg & R. S. Tindale, *Blackwell Handbook of Social Psychology: Group* Processes, Blackwell Publishers, 31-32

Strawson, G., 2003, "Mental Ballistics or the Involuntariness of Spontaneity," *Proceedings of the Aristotelian Society*, 103(3), 227-57.

Stel, M., van den Bos, K., & Bal, M., 2012, "On Mimicry and the Psychology of the Belief in a Just World: Imitating the Behavior of Others Reduces the Blaming of Innocent Victims," *Social Justice Research*, 25(1), 14-24.

Tetlock, P. E., 1983, "Accountability and the Perseverance of First Impressions," *Social Psychology Quarterly,* 46(4), 285-92.

Tetlock, P. E., 1985, "Accountability: A Social Check on the Fundamental Attribution Error," *Social Psychology Quarterly*, 48(3), 227-36.

Tetlock, P. E., & Kim, J. I., 1987, "Accountability and Judgement Processes in a Personality Prediction Task," *Journal of Personality and Social Psychology*, 57, 388-98.

Tetlock, P. E., Skitka, L., & Boettger, R., 1989, "Social and Cognitive Strategies for Coping with Accountability: Conformity, Complexity, and Bolstering," *Journal of Personality and Social Psychology, 57*(4), 632–640.

Thompson, D. F., 2008, "Deliberative Democratic Theory and Empirical Political Science." *Annual Review of Political Science, 11*, 497–520.

Tiberius, V., 2002. "Practical Reason and the Stability Standard." *Ethical Theory and Moral Practice*, 5, 339–53.

Tiberius, V., 2013, "In Defense of Reflection," *Philosophical Issues*, 23(1), 223-43.

Tolman, E., E., 1948, "Cognitive Maps in Rats and Men," *Psychological Review*, 55(4), 189-208.

Tversky, A., & Kahneman, D., 1981, "The Framing of Decision and the Psychology of Choice," *Science,* Vol. 211, No. 4481, 452-458.

Vinokur, A., & Burnstein, E., 1978, "Depolarization of Attitudes in Groups," *Journal of Personality and Social Psychology,* 36(8), 872-85.

Velleman, J.D., 2000a, *The Possibility of Practical Reason*, Clarendon Press.

Velleman, J. D., 2009, *How We Get Along*, Cambridge University Press.

Wagner, D., 2002, *The Illusion of Conscious Will*, MIT Press.

Wigboldus, D. H. J., Sherman, J. W., Franzese, H. L., & van Knippenber, A., 2004, "Capacity and Comprehension: Spontaneous Stereotyping Under Cognitive Load," *Social Cognition*, 22(3), 292-309.

Williams, B., 1976, "Persons, Character, and Morality," in A. O. Rorty (ed.), *The Identities of Persons*, University of California Press, 197-216.

Williams, B., 1979, "Internal and External Reasons," in R. Harrison (ed.), *Rational Action*, Cambridge University Press, 101-113.

Wilson, G., & Shpall, S., 2002, "Action," *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/action/

Wilson, T. D., Dunn, D. S., Kraft, D., & Lisle, D. J., 1989, "Introspection, Attitude Change, and Attitude-Behavior Consistency: The Disruptive Effects of Explaining Why We Feel the Way We Do," In L. Berkowitz (ed), *Advances in Experimental Social Psychology*, 22, 287-343.

Wilson, T. D., & Schooler, J. W., 1991, "Thinking Too Much: Introspection Can Reduce the Quality of Preferences and Decisions," *Journal of Personality and Social Psychology*, 60, 181-92.

Wheatley, T., & Haidt, J., 2005, "Hypnotic Disgust Makes Moral Judgments More Severe," *Psychology Science*, 16(10), 780-84.

Woodward, J., & Allman, J., 2007, "Moral Intuition: Its Neural Substrates and Normative Significance," *Journal of Physiology*, 101(4-6), 179-202.

Yarrow, K., Brown, P. & Krakauer, J. W., 2009, "Inside the Brain of an Elite Athlete: The Neural Processes that Support High Achievement in Sports," *Nature Reviews Neuroscience*, 10(8), 585-96.